AFOSR-91-0144: Final Project Summary

Title: "SPONTANEOUS DISCOVERY AND USE OF CATEGORICAL STRUCTURE"

This project aimed to investigate how human learners discover and learn about categories in the absence of explicit feedback from an external tutor (unsupervised learning). Most previous research on human category learning has investigated such learning as it occurs in supervised tasks, in which the experimenter is available to provide predefined categories and categorization-related feedback. Until recently, there were few empirical studies of unsupervised category learning, or even reliable methods for investigating such learning.

Our first goal was to develop reliable tasks and dependent measures for studying unsupervised learning in the laboratory. Given such tasks, we aimed to develop a general model or picture of the process by which unsupervised category learning occurs. An important step in this process is finding out whether the learning process is fundamentally accumulative and continuous, or whether there are important nonlinearities or discontinuities to be identified. One such discontinuity would be the discrete creation of new categories to describe novel or unusual stimuli that contradict the norms of existing categories. An important aim of our research has been to discover whether unsupervised learning proceeds through such all-or-none category invention, or whether learning occurs incrementally, as a result of gradually accumulating evidence about patterns of co-occurring properties across different training instances.

In addition to investigating the learning of general categories from a series of training instance, we also investigated how learning such categories affected the way in which individual training instances are encoded and remembered. These two issues are closely related, because if category learning affects how information is acquired about individual instances, then this in turn may influence the discovery and learning of further categories.

During the past three years of AFOSR funding, we have conducted experiments and pilot studies aimed at developing reliable empirical methods for investigating these issues, and applying these methods to distinguish among different theories of the learning process. To date, three articles have been written based on this research. The first was published in the March 1994 issue of the *Journal of Experimental Psychology: Learning, Memory, and Cognition*. The other two are included as part of the present report. One of these has been submitted for publication in the same journal, and the other will soon be submitted for publication pending final revisions. In addition, a major address was given overviewing this research in August 1994 at the *"Third Practical Aspects of Memory"* Conference held at the University of Maryland. A copy of that address is also enclosed.

We now provide a brief description of the three main articles.

1. Category Invention in Unsupervised Learning: describes three experiments using a free attribute-listing task as an index of unsupervised learning. This procedure allowed the learning of categories to be observed over trials, and levels of learning to be compared across different training conditions. The three experiments described in this article provided strong evidence that subjects invented categories in response to stimuli that violated the norms of previous categories. Learning was observed to vary not merely with the number of instances shown from a given category, but according to whether categories were presented such that each new category could be learned in contrast to the norms of a previous category. We refer to this dependence of learning on contrast rather than practice as a *contrast effect*, and it is such contrast effects that provide our primary evidence for category invention.

2. Instance and Category Learning in Unsupervised Tasks: describes five experiments in which a second task, based on memory for the features of individual training instances, was developed and tested as a procedure for investigating unsupervised learning. These experiments replicated the contrast effects obtained in the previous attribute listing experiments, and thus provided further evidence for the role of category invention in unsupervised learning. They also investigated learning of categories in which characteristic default features were present only probabilistically in individual training instances, as well as categories in which the default features were present in all instances. Evidence for learning by contrast-based category invention was obtained for both types of categories.

The task used in these experiments recorded how long subjects studied each feature of the training instances during the encoding phase of each trial, as well as their accuracy of verifying these features for subsequent recognition-memory tests. Thus, it allowed us to observe the effect of learning general categories on how subjects encoded and remembered individual instances. Evidence was obtained for an uncertainty-reducing encoding process based on selectively encoding those features of each instance not predictable from category norms ("schema-plus-corrections" encoding). These findings were consistent with prior studies of memory for text passages and other materials based on familiar categories (e.g., Bower, Black & Turner, 1979; Graesser, Woll, Kowalski & Smith, 1980). They also showed that category knowledge can improve learning of both expected and unexpected features of individual instances.

3. Category Invention and Transfer of Learning in Unsupervised Tasks: describes three experiments, two of which used the instance memorization task and one the attribute listing task to extend our earlier investigations of unsupervised learning into more complex stimulus domains. This research provided further evidence for the generality of category invention in discovering categories in non-feedback learning environments. Evidence was also obtained that the context of learning and the type of stimulus materials employed (pictorial vs. verbal stimuli) can affect the stability or confusibility of a set of categories after they have been learned. These experiments also provided evidence that new categories are learned in terms of their differences from similar existing categories, with shared features transferred to the new category rather being re-learned as if they were novel properties. This process by which new categories are created by adding minimal modifications or elaborations to existing knowledge is consistent with the schema-based encoding process used to learn individual instances. By conforming to this economizing principle, category invention provides an efficient means by which categories and subcategories can be acquired in complex stimulus domains.

To summarize our progress, in carrying out this project we have developed two new task paradigms for investigating unsupervised learning. We have obtained strong evidence for the use of category invention to distinguish different patterns of correlated features in these tasks. Such category invention is particularly interesting because it represents an important discontinuity or nonlinearity in the learning process, and because it suggests that learned contrast may often play a stronger role in discovering patterns than sheer exposure or practice. Importantly, the present tasks provide a way to study the acquisition of categories from individual training instances and at the same time investigate the effects of such category knowledge on how later instances are interpreted, processed, and remembered. The present results also suggest that new subcategories, as well as specific instances, are learned in terms of existing categories by encoding minimal modifications or specific details that set the new instance or category apart from the reference category in terms of which it is learned.

Unsupervised learning and pattern discovery are ubiquitous in everyday life, and are central components of many important performances in practical tasks. The major principles and underlying mechanisms of unsupervised learning have been poorly understood, and the topic has received little empirical study within experimental psychology. We argue that the present procedures and empirical results represent an important step in understanding these issues. We have studied the role of major factors such as practice, contrast, and the relation between categories in determining the course of category acquisition. Some of these factors, such as the importance of learned contrast in non-feedback environments, may have practical applications in the design of information systems and user interfaces. In addition, the task paradigms we have developed may be used in future studies of these and related issues in human learning.

SECURITY CLASSIFICATION OF THIS PAGE

Dist: A 8011

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited A

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 95 0007

6a. NAME OF PERFORMING ORGANIZATION
Dept. of Psychology

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
Same as 8A

6c. ADDRESS (City, State and ZIP Code)
Stanford University, Jordan Hall
Stanford, CA 94305

7b. ADDRESS (City, State and ZIP Code)
Same as 8C

8a. NAME OF FUNDING/SPONSORING ORGANIZATION Airforce Office of Scientific Research

8b. OFFICE SYMBOL (If applicable)
NL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-91-0144

8c. ADDRESS (City, State and ZIP Code)
Building 410
Bolling Air Force Base
Washington D.C. 20332-6448

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2313		

11. TITLE (Include Security Classification)
Spontaneous Discovery & Use of Categorical Structure

12. PERSONAL AUTHOR(S)
Gordon H. Bower & John P. Clapper

13a. TYPE OF REPORT
Technical /Final

13b. TIME COVERED
FROM 1/15/91 TO 1/14/94

14. DATE OF REPORT (Yr., Mo., Day)
94/12/15

15. PAGE COUNT
1 + 4 papers

16. SUPPLEMENTARY NOTATION
3 preprints enclosed

17. COSATI CODES

FIELD	GROUP	SUB. GR.
05	10	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Attention, concept, category, unsupervised learning

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The research project had as its goal the conduct of several experiments to examine people's ability to spontaneously classify and organize a large database of examples when no external tutor is there to inform them of the optimal organization. Throughout several experiments, we developed and tested three different, indirect measures of people's category learning. One set of those experiments led to a report published in the Journal of Experimental Psychology: Learning, Memory and Cognition. Copies of that published paper are enclosed. In addition, further experiments were conducted which yielded useful confirmatory results. These results have been written up and submitted for publication. The period of the grant extended without cost to October 15, 1994 to enable the writing of these papers. A summary of the papers is enclosed. This is the final report on the project.

DTIC QUALITY INSPECTED 3

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS □

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
John Tagney, Ph.D.

22b. TELEPHONE NUMBER (Include Area Code)
202-767-5021

22c. OFFICE SYMBOL
AFOSR/NL

EDITION OF 1 JAN 73 IS OBSOLETE.

unclassified
SECURITY CLASSIFICATION OF THIS PAGE

Journal of Experimental Psychology:
Learning, Memory, and Cognition
1994, Vol. 20, No. 2, 443-460

Category Invention in Unsupervised Learning

John P. Clapper and Gordon H. Bower

This research aimed to discriminate between 2 general approaches to unsupervised category learning, one based on learning explicit correlational rules or associations within a stimulus domain (autocorrelation) and the other based on inventing separate categories to capture the correlational structure of the domain (category invention). An "attribute-listing" paradigm was used to index unsupervised learning in 3 experiments. Each experiment manipulated the order in which instances from 2 different categories were presented and evaluated the effects of this manipulation in terms of the 2 competing theoretical approaches to unsupervised learning. Strong evidence was found for the use by Ss of a discrete category invention process to learn the categories in these experiments. These results also suggest that attribute listing may be a valuable method for future investigations of unsupervised category learning.

The study of concepts and category learning has long been a focus of research in cognitive psychology. Most of this research has studied supervised category learning, in which a tutor provides the subjects with category labels and feedback relevant to the success criterion of the learning task (e.g., Bruner, Goodnow, & Austin, 1956; see Millward, 1971, for a review). By contrast, unsupervised learning has received much less attention by experimental psychologists. In unsupervised learning, subjects must invent and use categories without predefined category labels or feedback from an external tutor. Many categories that people learn in real life are acquired in observational, untutored conditions and thus are examples of unsupervised learning. Much of our knowledge about the properties and behavior of common physical objects, social interactions, linguistic classes and rules, and everyday tasks and procedures may be learned in this manner (Billman & Heit, 1988). Any learning by pioneers about a novel environment is unsupervised because they must invent their own categories for describing that environment and generate their own criteria for classifying stimuli into these categories.

This article describes a recently developed procedure for investigating unsupervised learning (see Clapper & Bower, 1991) and three experiments in which this procedure was used to test theories of how categories could be learned and represented in unsupervised tasks. We begin by describing more precisely what we mean by an unsupervised learning task and how categories could be defined within such a task. We then argue that models of unsupervised learning can be divided into two general types, which differ in how category knowledge is represented in long-term memory and the processes by which this knowledge is abstracted from individual training instances. After providing this background, we describe the attribute-listing paradigm and show how it can be used to discriminate between the two classes of theories described earlier.

Defining Categories in Unsupervised Tasks

In supervised learning tasks, categories are predefined by the experimenter and subjects must use the experimenter's feedback to determine the correct rules for assigning stimuli to each category. Any arbitrary categorization rule may be used in such experiments (e.g., disjunctive rules such as "Members of Category A are either red squares or blue diamonds, but not red diamonds or blue squares"), and categories need not be functionally natural or capture informative patterns within the stimulus set. In contrast, in unsupervised tasks categories are not arbitrarily predefined by an external tutor; rather, subjects must discover categories for themselves as they explore a given stimulus domain. This presumably requires that some regularity or structure actually exist within that domain, that is, a pattern or signal that can be distinguished from the noise of background stimulus variation. It is necessary to define what kind of pattern or structure constitutes a category before proceeding to evaluate whether subjects in a given condition have learned this category.

Following Clapper and Bower (1991), we adopt a conventional feature-based vocabulary for describing commonalities and differences within a stimulus set and then define categories in terms of this vocabulary. Individual stimuli are described as collections of *features.* Each feature can be thought of as a specific, concrete *value* of a more generic or abstract *attribute.* For example, the stimuli in a given set could be described in terms of their shape (a generic attribute), with particular stimuli being squares, circles, or triangles (the specific values of the shape attribute). In principle, the values of an attribute could be either *discrete* (e.g., squares vs. circles) or *continuous* (i.e., ordered quantities, such as gradations of size or shading), but only the discrete-valued case is considered here.

Given a set of attributes for describing a stimulus domain, patterns of correlated features (attribute values) provide an inductive basis for partitioning that domain into subsets or

John P. Clapper and Gordon H. Bower, Department of Psychology, Stanford University.

This research was supported by U.S. Air Force Office of Scientific Research Grant AFOSR-91-0144.

Correspondence concerning this article should be addressed to John P. Clapper, Department of Psychology, Building 420, Stanford University, Stanford, California 94305.

Stimulus Set 1

Attribute									Attribute							
1	2	3	4	5	6	7	8		1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1		2	2	2	2	2	1	1	1
1	1	1	1	1	1	1	2		2	2	2	2	2	1	1	2
1	1	1	1	1	1	2	1		2	2	2	2	2	1	2	1
1	1	1	1	1	1	2	2		2	2	2	2	2	1	2	2
1	1	1	1	1	2	1	1		2	2	2	2	2	2	1	1
1	1	1	1	1	2	1	2		2	2	2	2	2	2	1	2
1	1	1	1	1	2	2	1		2	2	2	2	2	2	2	1
1	1	1	1	1	2	2	2		2	2	2	2	2	2	2	2

Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

Stimulus Set 2

Attribute									Attribute							
1	2	3	4	5	6	7	8		1	2	3	4	5	6	7	8
1	1	1	2	2	1	2	1		2	2	2	1	1	2	2	2
1	1	1	1	2	2	1	1		2	2	2	2	2	1	1	1
1	1	1	1	1	1	1	1		2	2	2	2	1	1	2	1
1	1	1	1	2	2	2	1		2	2	2	1	1	2	2	1
1	1	1	1	2	1	1	2		2	2	2	2	1	1	1	2
1	1	1	2	1	1	1	2		2	2	2	2	2	2	1	1
1	1	1	2	2	2	2	2		2	2	2	1	1	2	1	2
1	1	1	1	1	2	2	2		2	2	2	2	2	2	2	2

Category "A" : 1 1 1 x x x x x
Category "B" : 2 2 2 x x x x x

Stimulus Set 3

Attribute									Attribute							
1	2	3	4	5	6	7	8		1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1		3	2	3	2	4	4	3	4
1	1	1	1	1	1	1	2		4	3	3	4	2	4	3	3
1	1	1	1	1	1	2	1		2	4	4	4	2	3	4	3
1	1	1	1	1	1	2	2		4	3	2	2	4	3	3	3
1	1	1	1	1	2	1	1		3	2	4	3	4	4	4	4
1	1	1	1	1	2	1	2		2	4	3	3	3	4	4	3
1	1	1	1	1	2	2	1		4	3	2	3	3	3	3	4
1	1	1	1	1	2	2	2		3	3	4	2	3	3	4	4

Category "A" : 1 1 1 1 1 x x x
"Not - A" : y y y y y y y y

Stimulus Set 4

Attribute									Attribute							
1	2	3	4	5	6	7	8		1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1		2	2	2	1	2	1	1	1
1	1	1	2	1	1	1	2		1	2	2	2	2	1	1	2
1	1	1	1	1	1	2	1		2	2	2	2	2	1	2	1
1	2	1	1	1	1	2	2		2	2	2	2	2	1	2	2
1	1	1	1	1	2	1	1		2	2	2	2	2	2	1	1
1	1	1	1	1	2	1	2		2	1	2	2	2	2	1	2
1	1	1	1	2	2	2	1		2	2	2	2	2	2	2	1
1	1	1	1	1	2	2	2		2	2	2	2	2	2	2	2

Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

Figure 1. Sample stimulus sets illustrating how categories are defined in terms of correlated attribute values.

categories.[1] To illustrate, a collection of fruit flies bred in a geneticist's laboratory could be described in terms of attributes such as size, eye color, wing shape, leg length, and so on. If it was then observed that individuals with long wings also had red eyes, large size, and long legs, whereas those with short wings had white eyes, small size, and short legs, these patterns of feature co-occurrences would form an inductive basis for recognizing two distinct categories of fruit flies within that population (Clapper & Bower, 1991).

Figure 1 shows several stimulus sets with different types of correlational patterns that could serve as a basis for partitioning them into separate categories. Within each of these sets, some attributes have strongly correlated values whereas others do not. For example, in Stimulus Set 1, the first five attributes listed have perfectly correlated values whereas the last three attributes vary independently. We refer to correlated values as *default* values of the category to which they give rise. Attributes that are uncorrelated within a given category are referred to as *variable* attributes.

[1] Of course, the existence of such patterns depends on the particular set of attributes used to describe a given stimulus set. Thus, the same set of stimuli might be categorized differently with respect to different sets of attributes. In principle, the categorization of a given stimulus and the attributes used to describe it are somewhat mutable and dependent on the task context and the other stimuli with which it is contrasted. In practice, experimenters usually define a set of canonical attributes by which a stimulus set is generated and described, and this determines the normatively "correct" categorization of that set to which subjects' actual performance is compared. This is reasonable, and will predict performance accurately, so long as the attributes actually used by subjects to describe the experimental stimuli approximately correspond to those assumed by the experimenter.

Figure 1 also illustrates another point. namely, that the interfeature correlations need not be perfect for categories to be distinguished based on these correlations (see Stimulus Set 4). In principle, a category would have positive utility so long as some of its features could be predicted with greater-than-chance reliability. This is consistent with the arguments of Wittgenstein (1953), Rosch (1975), and others that natural categories are not defined in terms of necessary and sufficient features, but rather are often characterized by probabilistic features and fuzzy boundaries. Furthermore, defining category structure in terms of predictive utility (i.e., feature correlations) is consistent with the functional role of categories in making predictions, drawing inferences, and completing patterns based on partial information (e.g., Clapper & Bower, 1991; Holland, Holyoak, Nisbett, & Thagard, 1986; Schank, 1982).[2]

Theories of Unsupervised Learning

We distinguish two general approaches to capturing correlational patterns, each of which has been implemented by several models in the empirical literature. The first approach is to represent feature correlation patterns directly, for example, within a correlational matrix, rather than partitioning the domain into separate categories. We will refer to this as the *autocorrelation* approach because models of this type assume that learners monitor the strengths of association (correlation) between individual pairs of features. The only learning mechanism required by this theory would be a process for modifying correlational associations or rules. Such associations would be strengthed by repetition and weakened by decay, interference processes, or both. If some features within a stimulus set were consistently correlated in their appearance, their strengths of association would increase relative to those of uncorrelated values. Given such a correlational record in memory, subjects could fill in missing features of an incomplete pattern, distinguish correlated from uncorrelated features, and perform other such inferences normally associated with category-level knowledge. It is also important to note that this inferential power could be gained without any explicit categorization of the stimulus set.

There are two general types of autocorrelation theories. The first assumes that correlational associations between all presented features are strengthened simultaneously on each trial (e.g., J. A. Anderson, 1977; J. A. Anderson, Silverstein, Ritz, & Jones, 1977; McClelland & Rumelhart, 1985; Rumelhart, Hinton, & McClelland, 1986). We can refer to these models as *matrix autocorrelators* because memory is viewed as a matrix of interfeature correlations that are continually updated by new experiences. A specific example of this class would be the one-layered autoassociator model of J. A. Anderson (1977). The second type of such autocorrelation theories are the rule-sampling or hypothesis-testing theories, in which correlational hypotheses are tested sequentially (usually one per trial) against the observed features in each instance (e.g., Billman & Heit, 1988; Davis, 1985). These rules are strengthened by confirmation and may be weakened by disconfirmation on a given trial. The main difference between these theories and the matrix models is in whether all the interfeature correlations provided by an instance are strengthened simultaneously or sequentially and how many interfeature correlations are updated on each trial.

The second approach to capturing correlational patterns in a stimulus domain is to explicitly partition that domain into separate categories and store information about each category in separate data structures (e.g., schemas or prototypes). Within this approach, which we refer to as *category invention,* feature correlations are represented indirectly by (a) partitioning stimuli into explicit subsets or categories in accordance with correlational patterns and (b) accumulating summary norms separately for each category. These summary norms contain information about the expected features of individual instances. If only stimuli that contain a particular pattern of correlated features are assigned to a given category, then norms computed across this selected subset of instances will capture their correlational structure.

The major issue for the learner, according to this theory, is determining when and on what basis to create new categories. In many statistical clustering models of category learning, it is assumed that the learner will first scan an entire set of stimuli before computing the optimal classification scheme for that set (e.g., Fried & Holyoak, 1984; Michalski & Stepp, 1983). This assumption is generally unrealistic for human learners. Because of attentional limitations, people must examine stimuli one at a time and update relevant category norms in response to each. Rather than computing global classification schemes across whole stimulus sets, humans are more likely to be opportunistic categorizers, creating new categories as they are needed to accommodate novel stimuli that do not fit into existing categories (e.g., J. R. Anderson, 1991; Clapper & Bower, 1991; Holland et al., 1986; Schank, 1982). We refer to this as the *incremental learning* assumption.

Incremental learning implies that subjects attempt to categorize each presented stimulus and that summary knowledge about the category provides a framework within which instances are described and compared with normative expectations. Assuming that there is a good enough fit between an instance and a known category, the features of that instance

[2] Note that this definition of categories in terms of correlational patterns does not imply that all members of a given category must be more similar to each other than to any nonmember. For example, in Figure 1, Stimulus Set 2, the instance 11122222, which is a member of Category A, is more similar to instance 22222222, which is a member of Category B, than to fellow Category A instance 11111111. We define categories in terms of predictive utility rather than in terms of similarity, or family resemblance (e.g., Rosch & Mervis, 1975). As shown by Figure 1, in some domains there may be no categorization scheme in which all members are more similar to each other than to any nonmembers. Nevertheless, there may be useful structure to be captured in such domains (i.e., correlational patterns among some attributes of the stimuli). This definition, however, does not exclude the possibility that the most natural categories, those that are easiest to learn and use, may have members that are highly similar to each other and dissimilar to members of other categories. Thus, our definition of categories does not contradict the arguments of Rosch and Mervis (1975) and others that so-called basic level categories tend to exhibit family resemblance structures.

will be used to update the norms of its category, that is, the expectedness or subjective probability of the presented attribute values will be incremented in the category norms. But in some cases, an instance may fit poorly into even the closest available category. For example, it may be describable in terms of the attributes associated with that category, but it may also violate several of its default expectations (see, e.g., Schank, 1982). In these cases, a new category could be created to accommodate that stimulus. When later instances are presented similar to that which triggered the new category, they will also be assigned to this category.

The norms for a given category might be represented in several ways, including prototypes, schemas, scripts, frames, various networks, and production rules (e.g., J. R. Anderson, 1991; Holland et al., 1986; Kahneman & Miller, 1986; Minsky, 1975; Rumelhart & Ortony, 1977; Schank, 1982; Schank & Abelson, 1977). All of these approaches are capable of representing statistical summaries of the properties of instances within a category, that is, of resembling a subjective probability distribution for the occurrence of different values of each attribute. For the present purposes, the differences between these various methods of representing category norms are relatively unimportant, and they are de-emphasized throughout this article. The major claims of the category invention approach pertain not to details of how category norms are represented in memory but rather to the explicit separation of norms from different categories on the basis of their perceived contrast and to the selective assimilation of instances to these categories. It is this discrete partitioning of experience that generates the major predictions that are tested here.

Category Learning in Discrimination Tasks

The first step in testing these theories is developing an appropriate task or paradigm in which unsupervised learning can be reliably observed and investigated. In the absence of such tasks, little prior empirical study of unsupervised category learning has occurred. Previous supervised classification experiments provide little guidance toward developing unsupervised learning tasks. Traditionally, supervised learning has been measured by classification accuracy, where subjects classify presented instances into alternative categories provided by the experimenter (e.g., Bruner et al., 1956). Because subjects in unsupervised learning are not given predefined categories, classification accuracy obviously cannot be used to measure learning in these tasks.

If categories are defined in terms of correlational patterns within a domain of stimuli, then acquisition of such categories would be implied whenever the subjects' performance reveals their sensitivity to these patterns. One indication of such sensitivity would be if subjects in certain tasks responded differently to correlated attribute values than to uncorrelated values. One task we have investigated that has these properties is presented to subjects as an instance-discrimination (identification) task in which subjects are asked to learn to distinguish among a set of presented stimuli so that they can respond uniquely to each one. In learning to identify each individual instance of a set, subjects must first learn how that instance differs from the other stimuli presented during training. In other words, subjects must learn which features or combinations of features specify that instance's unique identity and must exclude all possible lures within the presented stimulus set.

If the subject's task is to memorize a collection of stimulus patterns, then their labor can be greatly reduced by noticing and taking advantage of redundancies among some of the features. These advantages can be illustrated with the task of memorizing the 16 stimulus patterns shown as rows in Stimulus Set 1 of Figure 1. Here Attributes 1 through 5 are redundant, with values of 1 in one cluster (Category A) and values of 2 in the other cluster (Category B). Although there are eight attributes, and potentially $2^8 = 256$ patterns in the sample space, the 16 patterns actually presented can be uniquely identified by their values on four different attributes—the last three and some (any) one of the first five. Rather than memorizing the configuration of eight bits per stimulus, the optimal learner could memorize the 16 stimuli by recording only four bits of information for each pattern—namely, the category (or any of the default values, each of which predicts the other four) and the values of the last three (unpredictable or nonredundant) attributes. It is also important to note that once the value of one of the default attributes is specified, the other four defaults are unnecessary for identifying a unique stimulus. This contrast suggests that subjects who have learned the subjective categories (or clusters of interfeature correlations) will treat default attributes differently from variable attributes as they try to memorize each instance.

The Attribute-Listing Task

The foregoing discussion suggests that if an observable index of feature weighting could be developed for instance-discrimination tasks, then such tasks might be used to investigate unsupervised learning. In the experiments described in Clapper and Bower (1991), subjects were presented with a series of instances and were asked to list those features that they considered most informative for distinguishing each instance from all those that they had seen previously in the series. Subjects were told to imagine that they would have to use their feature list at some later time to pick out the current instance from among a field of similar distractors in a multiple-choice recognition test. They were instructed to list only those features they would need to pick out the current stimulus in such a discrimination test and to omit features that they would not need even if these were physically very salient or prominent. As in Figure 1, the stimuli in these tasks were composed of several attributes, each with two or more alternative values. Categories in the stimulus sets were defined in terms of correlated attribute values.

Within the autocorrelational approach, the probability of listing a given attribute value should depend on how strongly it is correlated with other values of the current instance. Thus, learning in a given condition is defined as subjects' sensitivity to differences in the degree of correlation among different pairs of attribute values, that is, sensitivity to the fact that some values of an instance are mutually redundant and others are not. This sensitivity is measured in terms of differences in listing probability for correlated versus uncorrelated values (i.e., in terms of subjects' observed preference for listing variables rather than defaults).

The interpretation of the listing task in terms of category invention is similar to its interpretation in terms of autocorrelation. Here, the probability of listing a given value should be a function of its expectedness or probability of occurrence within the current reference category. Learning is defined as sensitivity to differences in expectedness between variables and defaults, again measured by differences in their probability of listing (i.e., by subjects' preference for listing variables over defaults).

By subtracting the proportion of defaults listed on a given trial from the proportion of variables listed, we may compute a quantitative index of learning for that trial. This index provides a way to compare the level of learning on different trials of an experiment; for example, if the preference measure is statistically greater on Trial $n + 1$ than on Trial n, then it can be inferred that some learning has occurred over that interval of trials. In experiments reported by Clapper and Bower (1991), such a bias in favor of listing uncorrelated variables evolved gradually over trials as successive instances were encountered and subjects learned their consistent properties.

Distinguishing the Theories

In the present article, we use this preference measure to compare learning under different experimental conditions, that is, to evaluate the effects of specific independent variables on unsupervised learning. To test the autocorrelation versus category invention theories described earlier, we looked for some variable that the theories would expect to have different effects on learning. We noted that the theories we considered differed in their predictions of how the particular sequencing of training instances from two categories would affect the rate at which categories are learned. Consequently, the experiments described in this article rely on such sequence manipulations to test the autocorrelation versus category invention theories.

We assume that learners update their category knowledge (by modifying existing categories or creating new ones) following the presentation of each new training instance. Given this incremental learning assumption, category invention should be highly sensitive to the order in which instances from different categories are presented during training. In particular, learning should be greater when categories are acquired one at a time (e.g., when Category A is well-learned prior to encountering any instances of Category B) than when instances of different categories are presented together from the start of training. In the latter (mixed) sequence, learners may simply lump both types of instances together into a single category, thus, failing to capture the correlational patterns in the stimulus set.

To understand these predicted sequence effects, imagine an experiment in which instances of Category A are presented for the first n trials, followed by an instance of Category B on Trial $n + 1$. Given this arrangement, we would then ask how the probability of creating a new Category B on Trial $n + 1$ would vary as a function of n. To answer this question, consider that any reasonable function for inducing category norms from a series of training instances should show some sensitivity to basic statistical parameters (e.g., sample size and variability) that greatly affect the reliability of its norms (generalizations). For example, people should be more confident in assigning grey as the default color of elephants after they have seen many elephants, all of which were grey, than if they have seen only one elephant, which happened to be grey. Applying this observation to the experimental situation described above, as successive instances of Category A are presented (i.e., as the value of n is increased), one can see that the consistent default attributes of that category should increase in their expectedness. As the learner's confidence in the Category A norms increases, so should the perceived contrast between these norms and the first instance of Category B, which violates several default values of Category A. Thus, the probability of creating a new category in response to the first instance of Category B should increase with the number (n) of prior instances of Category A.

This analysis implies that presenting the first instance of Category B following only a few instances of Category A should lead to a higher probability that the two types of instances will be assimilated to a single, overarching category. This would occur because the features of the Category B instance would be compared with a relatively weak set of norms for Category A; hence, the perceived contrast between these norms and the instance of Category B would be reduced. If both types of instances were assimilated to a single category, the learner would then simply average over the feature correlations within the A and B categories, so that the correlational information conditional upon the two categories would be lost. Because neither instances in Category A nor instances in Category B would contrast strongly with this aggregated category on subsequent trials (assuming both were presented in random order), subjects might have difficulty unlearning these overgeneralized norms and discovering the correct category-level discriminations.

Autocorrelation models do not possess the same inherent tendency toward sequence sensitivity shown by category invention models when incremental learning is assumed. For example, it is easy to imagine a basic autocorrelation model that simply adds to incremental frequency counts within a correlational matrix each time a new instance is encountered. In principle, such a model would be completely immune to sequence effects on final learning (i.e., the final count in the matrix would be the same regardless of the order in which instances were presented). Thus, the model suggests that learning that the presence of large wings predicts black eyes in some insects would not affect learning that in other insects the presence of small wings predicts white eyes.

Although sequence sensitivity is not implied by the autocorrelational approach, it is important to ask whether it is possible to develop plausible models within this approach that mimic the particular type of sequence sensitivity expected by category invention. Existing autocorrelation models do not display sequencing effects similar to those of category invention. For example, autocorrelation models developed within the connectionist framework generally predict sequence effects that are almost the opposite of those expected by category invention models (e.g., J. A. Anderson, 1977; Rumelhart et al., 1986). These models predict that correlational learning should be improved if instances of both categories are presented mixed together (e.g., in random alternation) from the beginning of training. Presenting a block of instances from Category B following an earlier block of instances from Category A causes massive forgetting of correlational associations learned during the Category A block, a phenomenon referred to as *catastrophic interference* (McCloskey & Cohen, 1989; Ratcliff, 1990). By contrast, category invention theory predicts better learning in a blocked condition than when instances are presented in a mixed sequence and expects no catastrophic interference between categories.[3]

Models of unsupervised learning based on serial hypothesis testing (e.g., Billman & Heit, 1988; Davis, 1985) also fail to reproduce the sequence effects expected by category invention. In such models, there is little reason to expect interference between different correlational patterns within either blocked or mixed sequences. For example, the rule "large wings implies black eyes" neither confirms nor disconfirms the rule "small wings implies white eyes," and there is no obvious reason why learning one should increase the difficulty of learning the other. Indeed, the focused-sampling assumptions of Billman and Heit (1988) seem more compatible with positive transfer across categories (at least if the categories differ by contrasting defaults on the same set of attributes).

To reproduce the sequence effects predicted by category invention, autocorrelation models would have to include a process that strongly reduces correlational learning when instances of two patterns are mixed together, but not when they are presented in separate blocks. For example, an autocorrelation model could assume that correlational learning is subject to associative interference, or fan effects, similar to those studied in experiments on paired-associate learning (e.g., Postman, 1971) and sentence memory (e.g., J. R. Anderson, 1976; 1983). Thus, learning an association between a particular pair of attribute values (e.g., large wings with black eyes) might interfere with learning associations between other values of the same attributes (e.g., small wings with white eyes). This autocorrelation-with-interference theory could accommodate some of the results predicted by the alternative, category invention theory. For example, the interference theory predicts that correlations within Category A would be learned more slowly if instances of Category A were interwoven in the training sequence with instances of Category B than if the instances of Category A were presented alone or prior to any instances of Category B.

The problem with such interference theories is that interference should occur between different correlational patterns regardless of whether training instances are presented in a blocked or mixed sequence, whereas category invention predicts interference only in mixed sequences. Thus, an interference theory expects prior learning of instances of Category A to impair correlational learning in a later block of instances of Category B (similar to the negative transfer between lists observed in many paired-associate learning experiments (e.g., Postman, 1971). The more instances of Category A that are presented prior to Category B, the greater the negative transfer and the slower should be the learning of Category B correlations. This contradicts the prediction of category invention that presenting more instances of Category A prior to Category B should increase the probability of creating a separate Category B and thus improve learning of both Category A and B defaults.

Such correlational interference would also imply that learning Category B in a blocked sequence would cause retroactive interference and reduce prior learning of Category A (e.g., Postman, 1971), although this effect need not be as strong as the catastrophic interference predicted by connectionist autoassociators. (At least, catastrophic retroactive interference is not generally observed in standard experiments on associative interference.) As noted previously, the category invention theory expects no interference across categories once separate categories have been formed.

In summary, such variations of the autocorrelation approach appear unable to mimic the particular pattern of sequence sensitivity expected by category invention theories. Thus, demonstrating superior learning and a lack of interference between categories in blocked training sequences would provide evidence for a nonincremental, contrast-based process of category invention.

Experiment 1

The aim of this experiment was to evaluate the attribute-listing task as an index of unsupervised learning and to test the predictions of the two theories concerning sequence effects. Subjects' listing of attributes was compared in three conditions. In the blocked condition, the stimuli were partitioned into two categories based on patterns of correlated attribute values. The training instances were blocked by categories (i.e., a series of instances from one category was presented followed by a series of instances from the other category). Following these two training blocks was a test, or transfer, block in which several instances of both categories were presented in random order. In the mixed condition, the same instances were presented as in the blocked condition, but instances of both categories were randomly interspersed in the training sequence rather than being grouped into separate blocks. In the control condition, all the attributes of the stimuli varied independently, so that none of the attributes were correlated and the stimulus set was not partitioned into distinct categories. The same final test block that was presented in the

[3] Most connectionist models that could be applied to unsupervised learning are apparently subject to catastrophic interference, even when these models are not strict autocorrelators (but see Carpenter & Grossberg, 1987). The reason is that such models encode knowledge about contrasting categories as patterns of activation over the same set of network units even when these models do contain an explicit category level of representation (e.g., output units corresponding to different response categories). Because most connectionist models do not separate knowledge about different categories in memory the way that prototype or schema models do, different patterns are liable to interfere with each other, especially when they are learned separately, for example, in blocked training sequences (McCloskey & Cohen, 1989).

blocked condition was also given in the mixed and control conditions.

The first two conditions provided a test of the two models of unsupervised learning described earlier. Category invention implies that early aggregation may occur when contrasting categories are presented in a mixed sequence, and so poorer learning was predicted in the mixed condition than in the blocked condition. An autocorrelation model could accommodate interference between categories in the mixed condition by assuming that associative interference results from learning correlations among different values of the same set of attributes. However, this leads to the prediction that interference should be observed between the categories in the blocked condition as well as in the mixed condition, as noted earlier. Specifically, the autocorrelation-plus-interference hypothesis predicts (a) that the second category in the blocked condition should be learned more slowly than the first because of proactive interference or negative transfer from the first category, and (b) that once this second category is learned, it should produce retroactive interference on subjects' memory for the first category, that is, that evidence of forgetting or unlearning should be obtained when instances of the first category are presented in the final test block. By contrast, the category invention theory expects little interference of any kind in the blocked condition.

The third condition was included in this experiment as a control to evaluate learning in the other two conditions. This condition was identical to the others except that the stimuli lacked correlated attributes. Thus, any differences in performance between this condition and the correlated-attribute conditions would be due to these correlations rather than to other, extraneous, factors.

Method

Subjects. The subjects were 30 Stanford University undergraduates participating in partial fulfillment of an introductory psychology course requirement.

Procedure. Subjects were tested in groups of 8 to 10 for a single session of 40 to 50 min. The training instances were realistic line drawings of fictitious insects (see Figure 2) presented in a 42-page booklet that measured 8 in. by 11.5 in. (20.3 cm by 29.2 cm). The first two pages of this booklet contained full instructions and an agreement that subjects signed to indicate their informed consent to participate. A single training instance (insect picture) appeared on each subsequent page, together with brief instructions for the experimental task.

Subjects were instructed to write on each page the "distinctive" properties of each individual insect, where distinctive properties were those that would be useful for distinguishing the current instance from others of the same general type. Subjects were told to imagine that they were writing their lists in order to prepare for a later multiple-choice recognition test in which they would have to match up each list with the correct insect from among a large number of distractor items (i.e., other bugs from the same test booklet). Subjects were instructed to list only those properties that would be useful for identifying an insect on such a test and to omit nondistinguishing properties even if they were highly prominent or noticeable. They were further told to look only at the page of the booklet that they were currently working on and not to look backward or forward at other pages.

Subjects were allowed to complete the experimental task at their own pace. Once they had finished, they were given a debriefing page that explained the procedures and goals of the experiment and were allowed to leave.

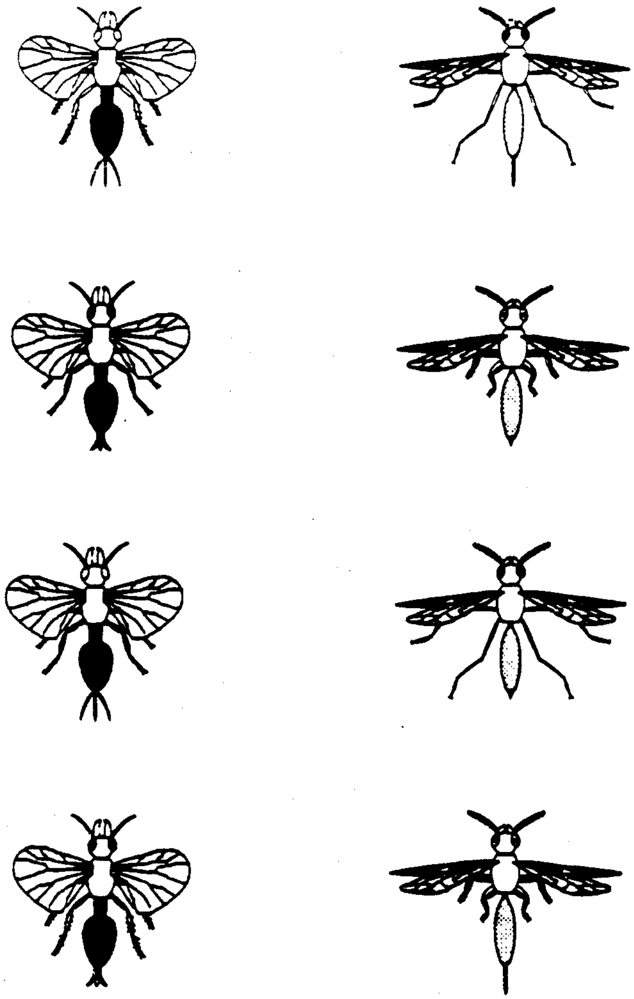


Figure 2. Sample stimuli from Experiment 1. Instances of one category are on the right and instances of the other are on the left. The correlated attributes in this stimulus set are wings, abdomen shape, abdomen shading, mandibles, and antennae; the variable attributes are legs, tails, and eyes.

Materials. The stimuli were line drawings of fictitious insects, all of which shared a common base structure (e.g., head, thorax, abdomen) plus eight dimensions of variation (attributes), such as wing shape, abdominal markings, eye color, and so forth (see Figure 2). Each attribute had either two or four discrete values (e.g., different wing shapes, differently colored eyes) depending on the experimental condition to which it was assigned.

The stimuli shown to a given subject were constructed according to one of two different plans depending on a subject's assigned condition (see Table 1). In two of the three experimental groups, the stimulus set was partitioned into two distinct categories defined by contrasting sets of correlated attribute values. In these correlated groups, five of the eight attributes were binary (two-valued), and their values were perfectly correlated across the instances such that each instance contained one of two possible sets of correlated values (denoted as Values 1 or 2 in Table 1). An instance's category membership was defined by which of these two clusters of correlated values it contained. These values are referred to as the *default* values of each category.

Table 1
Stimulus Set Design and Counterbalancing in Experiment 1

	Stimulus set							
	Correlated Set 1		Control Set 1		Correlated Set 2		Control Set 2	
Attribute	Category A	Category B	Category A	Category B	Category A	Category B	Category A	Category B
1. Wings	1	2	1	2	1	2	1	2
2. Body	1	2	1	2	1	2	1	2
3. Markings	1	2	1, 2	1, 2	1, 2	3, 4	1, 2	3, 4
4. Tails	1	2	1, 2	1, 2	1, 2	3, 4	1, 2	3, 4
5. Eyes	1	2	1, 2	1, 2	1, 2	3, 4	1, 2	3, 4
6. Legs	1, 2	3, 4	1, 2	3, 4	1	2	1, 2	1, 2
7. Jaws	1, 2	3, 4	1, 2	3, 4	1	2	1, 2	1, 2
8. Antennae	1, 2	3, 4	1, 2	3, 4	1	2	1, 2	1, 2

Note. The table entries indicate possible values of each attribute within a given stimulus set. The two correlated stimulus sets were each shown to two different groups of subjects, one in a blocked sequence and one in a mixed sequence.

The remaining three attributes in the correlated conditions had four values and were variable within each category. Two of the four values occurred with equal probability in instances of Category A, whereas the other two occurred with equal probability in instances of Category B. These attributes were uncorrelated within each category (i.e., they varied independently across instances of that category). Within these constraints, $2^3 = 8$ instances were generated from each category, for a total of 16 overall.

The stimuli in the control condition were equivalent to those in the two correlated conditions in the number of values assigned to each attribute (two or four), but these insects lacked correlated attributes present in the other conditions. Two attributes were correlated in all conditions; these were the wing shape and body shape attributes, which we judged to be the most salient attributes of the insects. These defaults, which were constant across all three groups, are referred to as *base defaults.* The four-valued variables were coordinated with the base defaults in the same way in the uncorrelated group as in the correlated groups (see Table 1). The stimuli in the uncorrelated groups can be divided into two pseudocategories on the basis of the base defaults and the pattern of dependent variation of the four-valued variables. However, three binary attributes that had correlated defaults in the other conditions occurred as uncorrelated variables in this condition.

The control condition was designed to show that any greater listing of variables over defaults in the correlated conditions could not simply be explained as an artifact due to variables possessing more possible values than defaults (four versus two). If this artifactual explanation is correct, then the same degree of bias in reporting variables over defaults should be observed in the control group as in the correlated conditions. But if the preference for listing variables over defaults is greater in the correlated groups than among the controls, this difference must be due to subjects' explicit or implicit correlational learning.

Design. The experimental design contained three between-subjects conditions, two of which had correlated values and one of which did not, as explained previously. The two correlated conditions used the same stimuli and differed only in the order in which training instances from the two categories were presented.

In the blocked condition, instances of Category A were presented in random order for the first 16 trials, followed by 16 trials in which instances of Category B were presented (each instance of the two categories was presented twice). Following this training phase was a final test block of eight trials in which four instances from each category were presented together in a mixed sequence. The order of instances in this test block was randomized (the same randomizations were used for subjects in all three groups), with the restriction that no more than two instances from the same category could occur in a row.

In the mixed condition, the same instances were presented as in the blocked condition, but in a different order. During the training phase, 16 instances from Category A and the 16 instances from Category B were presented in an intermixed sequence rather than blocked as in the previous condition. Instances from the two categories were presented in random order, with the restriction that no more than three instances from the same category could occur consecutively. A final mixed test block of eight instances from the two categories was then presented, the same as that used in the blocked condition, (i.e., the same specific insect pictures were presented in the same order in both conditions).

In the control condition, instances were presented in random order for the first 32 trials, except that no more than three instances with the same base default values were allowed to occur in a row during this phase. The final eight test trials were identical to those of the category conditions (i.e., five attributes were correlated during this block).

Counterbalancing the design. To construct stimuli from the specifications shown in Table 1, we first assigned particular stimulus attributes to abstract roles in the design. This assignment was held constant across all groups. With the exception of base defaults, each attribute had four values in half of the groups and two values in the other half of the groups. Two different stimulus sets were constructed for each of the three between-subjects conditions (blocked, mixed, and control); that is, six booklets were constructed and presented to different subjects. Attributes that were four-valued variables in one group were two-valued defaults in the other group from the same condition. This ensured that any effects due to materials (e.g., differences in the baseline salience or prominence of different attributes) would be balanced over the experiment as a whole.

Results and Discussion

Subjects' attribute lists were coded in terms of whether or not each of the eight relevant attribute dimensions was mentioned on a given trial.[4] The main index of learning was the

[4] Because this was a free-listing task, subjects generated their own response categories. For example, subjects shown a bug with large mandibles might describe the instance as possessing "big pincers," "large mouthparts," "oversized mandibles," or a variety of other labels. Although the specific labels might vary among different subjects, it was generally clear which attribute was being referred to at

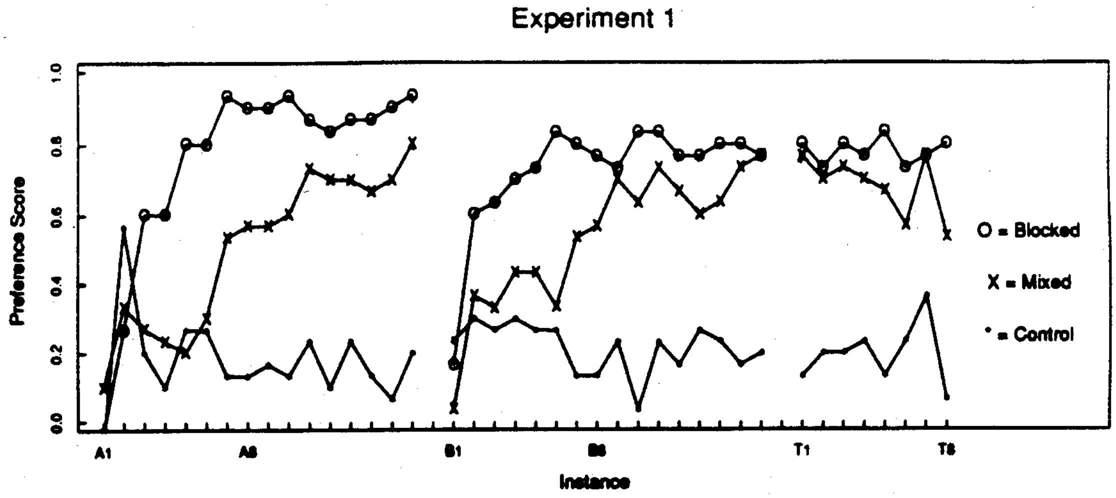


Figure 3. Preference scores for the three conditions of Experiment 1. Data from the mixed and control conditions are separated by category (or pseudocategories defined by the base defaults in the control condition), whereas trials from the test block are presented in their original order for all three groups.

proportion of variables listed minus the proportion of defaults listed on a given trial. This difference is referred to as the *preference score* for each trial because it reflects subjects' preference for listing variables over defaults. Preference scores for all three conditions are shown in Figure 3.

We also recorded the proportion of base defaults listed on each trial, but because they were correlated in all groups and hence potentially contaminated by materials effects, subjects' listing of these attributes did not provide the same unambiguous measure of learning as did the preference scores. Hence, we focus on preference scores as the principle dependent measure in most of the following discussion.

Examination of Figure 3 reveals, first, that preference scores were higher overall in the two correlated conditions (blocked and mixed) than in the control condition. Averaged over all trials, four-valued attributes were listed 19.6% more often than two-valued attributes in the control condition, a significant preference, $t(9) = 3.93$, $SE = 0.05$, $p < .01$. However, this preference was much stronger in the other two groups. Variable attributes were listed 74% more often than defaults in the blocked condition, $t(9) = 9.65$, $SE = 0.077$, $p < .001$. This preference was significantly greater than the corresponding difference in the control condition, $t(18) = 5.95$, $SE = 0.092$, $p < .001$. In the mixed condition, variables were listed about 55% more often than defaults, $t(9) = 10.38$, $SE = 0.053$, $p < .001$; this effect was also significantly larger than the 19.6% preference in the control condition, $t(18) = 4.84$, $SE = 0.073$, $p < .001$.

The contrasting results for the correlated versus control conditions indicate that the preference for listing variables over defaults in the correlated conditions was in large part due to the correlations themselves, not simply to the fact that uncorrelated attributes had a larger number of possible values than correlated attributes. Thus, subjects in the correlated groups must have internalized the correlational structure of the stimulus set in some manner, either by tracking pairwise correlations or by partitioning the set into separate categories.

The category invention theory predicts that preference scores would show rapid learning of both categories in the blocked condition but that learning in the mixed condition would be slower. The data are generally consistent with this prediction. If one examines the data plotted in Figure 3, it is apparent that the preference for studying variables over defaults increased rapidly for both categories in the blocked condition. Preference scores increased from −.03 on the first Category A trial to .90 on the eighth and remained fairly stable thereafter; the linear trend over the first eight trials was highly significant, $t(9) = 9.01$, $SE = 0.59$, $p < .001$. Subjects sharply increased their listing of defaults when the first instance of Category B was presented. The resulting decrease in preference scores, compared with the immediately preceding Category A trial, was highly significant, $t(9) = 6.31$, $SE = 0.122$, $p < .001$. Thereafter, preference scores increased rapidly from .17 on this first Category B trial to a maximum of .83 by the sixth. The linear contrast over the first half of this block was statistically significant, $t(9) = 4.58$, $SE = 0.63$, $p < .01$. However, no significant change occurred over the remaining nine instances in this block.

Recall that the autocorrelation-with-interference hypothesis predicts that prior learning of Category A should reduce subsequent learning of Category B because of negative transfer or proactive interference effects. But no such interference occurred in the present experiment. Learning of Category B appeared to occur at least as rapidly as that of Category A, and there was no significant difference between asymptotic learning of the two categories (i.e., when preference scores averaged over the last eight instances of each were compared), $t(9) = 1.54$, $SE = 0.065$, $p > .10$. This absence of proactive

any given time. In the few cases in which there was some initial uncertainty about which attribute a subject meant to refer to, this was resolved by observing how the use of the label shifted over the next few trials in correspondence with changes in the particular attributes under consideration.

interference appears to be a strike against the autocorrelation theory but is consistent with category invention models.

Preference scores during the mixed test block did not differ significantly from those of the earlier blocks. This was true when the test block was compared to the last eight instances of Category A, $t(9) = 1.64$, $SE = 0.067$, $p > .10$, as well as to the last eight instances of Category B, $t(9) = 0.35$, $SE = 0.022$, $p > .50$. In addition, preference scores during the test block did not differ between the two categories, $t(9) = 0.16$, $SE = 0.504$, $p > .50$. These results indicate that the learning observed during the earlier training blocks, in which instances of the same category were presented for many trials in succession, generalized to a different context in which the two categories were mixed. In other words, learning was stable over changes in the learning environment (Carpenter & Grossberg, 1987). There was no evidence for retroactive interference from learning Category B upon test performance on Category A, as would have been expected in an autocorrelation-with-interference framework.

Learning occurred somewhat more slowly in the mixed than the blocked condition. Preference scores increased over the entire training block for each category. This increase was significant for both Category A, $t(9) = 4.18$, $SE = 3.55$, $p < .01$ and Category B, $t(9) = 3.74$, $SE = 3.31$, $p < .01$. However, preference scores were greater in the blocked than the mixed condition over the first eight instances shown of Category A, $t(18) = 3.69$, $SE = 0.079$, $p < .01$ and of Category B $t(18) = 2.34$, $SE = 0.12$, $p < .05$. The same comparison was marginally significant over the second eight instances of Category A, $t(18) = 1.96$, $SE = 0.10$, $p < .10$. Pooled over all 32 training trials, preference scores were significantly higher in the blocked than the mixed condition, $t(18) = 2.46$, $SE = 0.089$, $p < .05$. However, the blocked and mixed conditions did not differ significantly during the test block, $t(18) = 0.71$, $SE = 0.14$, $p > .25$. This suggests that although learning occurred more rapidly in the blocked condition, subjects in the mixed condition were able to catch up by the end of training.

The faster learning that was due to category blocking is consistent with the category invention theory because it expects subjects to have difficulty separating categories presented in a mixed sequence. As noted earlier, however, an autocorrelation model could explain negative transfer in the mixed condition as being due to interference or unlearning of correlations among different feature pairs. However, such an interference process predicts a different pattern of results in the blocked condition than were shown by these data. First, it implies that prior learning of Category A should interfere with subsequent learning of Category B. However, these data show no such negative transfer; the second category was learned at least as fast as the first in this group. Second, an autocorrelation-with-interference model also predicts that Category B should exert strong retroactive interference on Category A in the blocked condition. As noted earlier, no evidence of such interference was obtained in the final test trials of this experiment. This lack of retroactive interference is particularly embarrassing for connectionist autocorrelators, which predict catastrophic interference from learning the Category B correlations on subjects' memory for the earlier Category A correlations (McCloskey & Cohen, 1989; Ratcliff, 1990).

Although the preference scores showed no evidence of retroactive interference during the test block, there was some evidence that presenting instances of the two categories in a mixed sequence increased the salience of their category membership. Recall that base defaults were the most physically prominent attributes of the insect stimuli, and it was considered likely that subjects would tend to list these particular attributes when they wished to indicate an instance's category membership. Although caution must be exercised when interpreting listing patterns for base defaults, because these attributes were correlated in all groups and hence their data may be contaminated with unbalanced materials effects, it appears that base defaults were often used by subjects to indicate the categorization of each instance. Consistent with this explanation, higher listings were observed for base defaults in the mixed test block of the blocked condition, in which the categorization of instances varied from trial to trial, than in the last eight trials of the preceding same-category training blocks, in which categories were constant and could be inferred from local context, $t(9) = 2.48$, $SE = 0.081$, $p < .05$. No such increase occurred for either variables, $t(9) = 1.00$, $SE = 0.004$, $p > .25$, or for regular defaults, $t(9) = 1.54$, $SE = 0.035$, $p > .10$. In other respects, the base defaults behaved like the regular defaults in the blocked condition, decreasing strongly during the first six instances of each category: $t(9) = 2.83$, $SE = 0.478$, $p < .05$ for Category A, and $t(9) = 6.85$, $SE = 0.255$, $p < .001$ for Category B.

By contrast, base defaults remained fairly constant throughout the experiment in the mixed condition, showing no significant decreasing trends and remaining significantly higher than the regular defaults, $t(9) = 2.55$, $SE = 0.067$, $p < .05$. Subjects who learned the categories in the mixed condition would have needed to explicitly indicate the category membership of each instance throughout the experiment because this could not be inferred from context. To do so, they should have continued listing at least one of the base defaults as shown by the present data.

Experiment 2

The aim of this experiment was to extend the results of Experiment 1 by testing further predictions of the category invention theory. Subjects were randomly assigned to two conditions. In the contrast condition, a pretraining block of 8 instances of Category A was followed by a test block of 12 instances of Category A and 12 instances of Category B that were presented in mixed sequence. In this condition, subjects should learn strong Category A defaults prior to encountering their first instance of Category B. They should readily notice the contrast between the two categories when they encounter this instance of Category B and rapidly learn the default values of the newly invented Category B without unlearning or weakening the prior Category A norms.

In the second, practice, condition, a mixed pretraining block of four instances of Category A and four instances of Category B was followed by the same test block as in the contrast condition. Category invention implies that subjects may aggregate the two types of instances into a single category, thereby pooling and obscuring the correlational structure of the stimulus set. The result would be reduced learning of both categories in this condition. By contrast, the autocorrelation theory expects better learning of Category B in the practice condition because correlational associations among Category B defaults would receive more practice (repetitions across different instances) in that condition. (A total of four instances of Category B were presented during the pretraining block in the practice condition, whereas no instances of Category B occurred prior to the test block in the contrast condition.)

The theories also make different predictions about transfer of learning from one category to the other. First, increasing the number of instances of Category A, from four in the practice condition to eight in the contrast condition, is expected by category invention theorists to improve later learning of Category B. This would seem to be an example of positive transfer from Category A to Category B. Second, increasing the number of instances of Category B, from zero in the contrast condition to four in the practice condition, is expected by category invention theorists to impair learning of Category A, which is an example of interference or negative transfer from Category B to Category A. This seeming paradox—positive transfer from A to B combined with negative transfer from B to A—makes sense in terms of category invention because this theory assumes that the particular sequence in which instances are presented affects the probability of creating separate categories by either highlighting or camouflaging the differences between them. By contrast, the predicted interaction of transfer and repetition effects with the sequencing and number of instances from each category makes little sense within the autocorrelational framework. If the predicted pattern of results obtains, it would provide strong evidence for the existence of a category invention process in unsupervised learning.

Method

Subjects. The subjects were 40 undergraduate students of San Jose State University participating in partial fulfillment of an introductory psychology course requirement.[5]

Procedure. Subjects were tested in groups for a single session of 30–45 min. The training instances were line drawings of fictitious insects presented in booklets similar to those used in Experiment 1. The attribute-listing procedure was identical to that of Experiment 1, except that the present experiment consisted of 32 instead of 40 trials.

Materials. The same type of pictorial insect stimuli as in Experiment 1 were used. These stimuli all shared a common base structure (e.g., head, thorax, abdomen) plus eight dimensions of variation (attributes), such as wing shape, abdominal markings, eye color, and so forth. Five of the eight attributes had two values, and these values were correlated across instances such that the set was partitioned into two distinct categories defined by contrasting sets of default attribute values (see Table 2).

The remaining three attributes had four values, two of which occurred with equal probability in Category A and the other two of which occurred with equal probability in instances of Category B. These variable attributes were uncorrelated within each category (i.e., they varied independently across instances of that category). A total of eight instances (2^3) could be generated within each of the two categories within these constraints. All 16 possible instances were presented to subjects in this experiment.

Design. Two between-subjects conditions were tested in this experiment. In the contrast condition, only instances of Category A were presented for the first eight trials, followed by a mixed block of 12 instances of Category A and 12 instances of Category B. The first block of eight trials was referred to as the *pretraining* block, whereas the second block of 24 trials was referred to as the *test* block. The first instance of the test block was always a member of Category B. Instances of both categories were thereafter presented in a randomly ordered, intermixed sequence, with the constraint that no more than three instances from the same category be allowed to appear in succession.

In the practice condition, the eight instances from the pretraining block consisted of four from Category A and four from Category B, rather than eight from Category A as before. The four instances from each category were selected so that both values of each variable attribute occurred twice, and none of the variable attributes was correlated with any of the others. These instances were presented in a random order, with the restrictions that the first instance be a member of Category A and that no more than two instances from the same category occur in sequence. The same 24-instance test block was used as in the contrast condition. Note that the only difference between the two conditions is that in the practice condition, four instances of Category B were substituted for the four instances of Category A that were presented in the contrast condition.

Counterbalancing the design. The counterbalancing scheme for this experiment is illustrated in Table 2. All of the attributes had four values in one condition and two (correlated) values in the other, except for the first two attributes. The first two attributes were base defaults, which consisted of the wing shape and body shape attributes, as in Experiment 1. These were two-valued and correlated in both conditions. The balancing scheme shown in Table 2 ensured that materials effects (e.g., differences in baseline prominence of different attributes) would be balanced over the six attributes that were not base defaults. Half of the subjects in the contrast and practice conditions were tested with Stimulus Set 1 and half with Stimulus Set 2.

Table 2
Stimulus Set Design and Counterbalancing in Experiments 2 and 3

	Stimulus set			
	Group 1		Group 2	
Attribute	Category A	Category B	Category A	Category B
1. Wings	1	2	1	2
2. Body	1	2	1	2
3. Markings	1	2	1, 2	3, 4
4. Tails	1	2	1, 2	3, 4
5. Eyes	1	2	1, 2	3, 4
6. Legs	1, 2	3, 4	1	2
7. Jaws	1, 2	3, 4	1	2
8. Antennae	1, 2	3, 4	1	2

Note. The table entries indicate possible values of each attribute within a stimulus set.

Results and Discussion

The same attribute-listing data as in Experiment 1 was collected in this experiment. The preference scores (listing of

[5] Thanks to Forest Jourden and to the San Jose State University Psychology Department for facilitating access to their subject pool for this experiment.

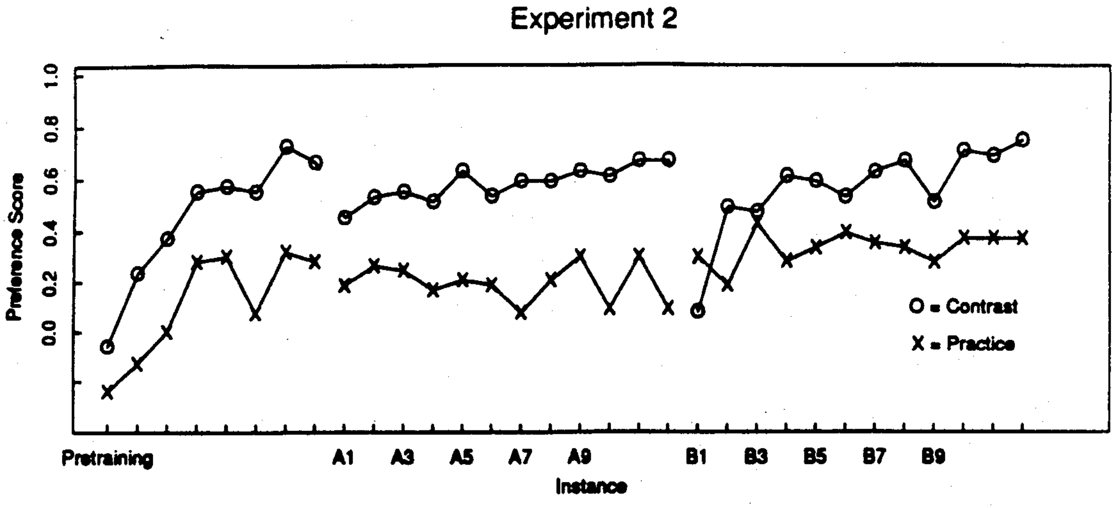


Figure 4. Preference scores for the two conditions of Experiment 2. Pretraining trials are shown in their original order, whereas the test block trials are separated by category.

variables minus that of defaults on each trial) displayed in Figure 4 were the main index of learning.[6]

As shown in Figure 4, learning was higher in the contrast than in the practice condition throughout the experiment. Preference scores increased significantly during the pretraining block in both contrast, $t(16) = 5.23$, $SE = 0.774$, $p < .001$, and practice conditions, $t(17) = 4.86$, $SE = 0.627$, $p < .001$; however, preference was higher overall in the blocked condition, $t(33) = 4.83$, $SE = 0.707$, $p < .001$. This result is not surprising because subjects in the mixed condition were shown instances of two categories during pretraining while those in the blocked condition only had one category to learn during this interval.

Turning to the test block (the numbered trials in Figure 4), one can see that learning of both categories was higher in the contrast condition than in the practice condition. Preference scores for Category A showed a significant decrease on the first Category A instance of the test block, relative to the last instance of the pretraining block, $t(16) = 2.85$, $SE = 0.075$, $p < .02$. Thus, encountering the first instance of Category B at the beginning of the test block appeared to have a significant effect on Category A norms in this group. After this initial decrease, preference scores for Category A showed a modest but statistically significant increase over the remaining trials of the test block, $t(16) = 2.36$, $SE = 0.959$, $p < .05$. By contrast, preference scores for Category A in the practice condition showed neither the initial decrease, $t(17) = 0.96$, $SE = 0.096$, $p > .50$, nor the subsequent increasing trend, $t(17) = -0.78$, $SE = 0.762$, $p > .20$, observed in the contrast condition. Overall, preference scores for Category A during the test block were higher in the contrast condition than in the practice condition, $t(33) = 3.40$, $SE = 0.114$, $p < .01$.

Note that both the category invention and autocorrelational approaches can accommodate the finding that overall learning of Category A was greater in the contrast than in the practice condition. Such a result is expected in category invention theory because subjects in the contrast condition had the opportunity to learn Category A in isolation, with no danger of initially lumping it with Category B, as occurred in the practice condition. Autocorrelation theory would expect better learning of Category A in the contrast condition because a larger number of instances in that category were presented in that condition.

Although both theories predict faster learning of Category A in the contrast condition, the autocorrelation approach has difficulty accommodating the detailed pattern of results from this condition. Thus, autocorrelation seems to imply that learning of Category A should have continued to increase following the pretraining block in the practice condition. Although Category A learning in the practice condition would have been expected to lag a few trials behind that in the contrast condition, in principle, asymptotic learning should have been about the same in both groups. However, Category A preference scores did not increase further during the test block of the practice condition; Category A learning appears to have stopped by the end of pretraining and never to have approached the asymptotic level attained in the contrast condition.

Category invention theory predicts that subjects should discriminate between contrasting categories better when one of the categories is learned first (contrast condition) than when instances of both are presented together from the start of training (practice condition). It is important to note that this result was predicted not only for the pretrained category (A) but also for the nonpretrained category (B). The better initial learning of Category A in the contrast condition was expected

[6] Five subjects were excluded from the data analysis, 2 from the practice condition and 3 from the contrast condition, because they produced no usable data from more than one third of the 32 trials in the experiment. A subject was considered to have produced no usable data from a given trial if he or she listed no features on that trial (i.e., left that page in the booklet blank), if the only information provided was a comparison to a previous instance (e.g., "same as the first one"), or if none of the features listed were representable within our eight-attribute coding scheme.

to increase the perceived contrast between the Category A norms and the features of the first instance of Category B, thereby increasing the probability that a new category would be created to describe the instance of Category B. Thus, better learning of Category B was predicted to occur in the contrast condition despite the larger number of instances presented to subjects in the practice condition (i.e., in spite of the fact that the interfeature correlations of Category B would have been repeated for a larger number of trials in the practice condition).

Consistent with this prediction, Category B was learned significantly better in the contrast condition than in the practice condition. Preference scores for this category increased quite rapidly in the contrast condition; the linear trend computed over the 12 trials of the test block was significant, $t(16) = 4.36, SE = 1.210, p < .001$. The linear contrast over the first 12 instances of Category B of the practice condition (4 of which were in the pretraining block) also showed a significant increase, $t(17) = 2.63, SE = 1.269, p < .02$. However, overall learning of Category B was higher in the contrast condition than in the practice condition. Averaged over the 12 test trials, preference scores in the contrast condition were significantly higher than those in the practice condition, $t(33) = 2.09, SE = 0.109, p < .05$. When asymptotic learning was compared by averaging the last six Category B trials in each condition, preference scores averaged 31.2% higher in the contrast condition, $t(33) = 2.89, SE = 0.108, p < .01$.

Although more instances of Category B were presented in the practice condition than in the contrast condition, subjects in the contrast condition showed greater learning of Category B. Presenting four instances of Category B during the pretraining block of the practice condition strongly interfered with the later learning of both categories in that condition. According to the category invention theory, this interference was due to inadequate learning of Category A defaults prior to encountering the first instance of Category B, which caused subjects to aggregate both types of instances into a single category.

In summary, the results of the present experiment were consistent with category invention and cannot be accommodated easily within a strictly autocorrelational approach. The only qualification of this support for category invention derives from the temporary increase in the listing of Category A defaults that occurred after the first instance of Category B was presented in the contrast condition. There are several ways to interpret this slight readjustment of Category A norms at the start of the test block. In theory, the instance of Category B should have triggered the invention of a new category and thus have had no effect on Category A norms nor on preference scores for subsequent instances of Category A. One possibility is that the first instance of Category B triggered a new category as expected, but that the instance was assimilated both to this new category and to Category A. The new category would then provide a better match to subsequent instances of Category B than would Category A, so for these later instances only the new Category B would be evoked. Meanwhile, the Category A norms would gradually return to previous levels as subsequent instances of Category A were assimilated and overwhelmed the effects of the earlier Category B values.

A related possibility (L. W. Barsalou, personal communication, October 14, 1992) is that the temporary reduction in Category A norms might have been due to subjects' explicitly contrasting the two categories during the early portion of the test block. People sometimes characterize categories in terms of their contrast with neighboring categories, as *male* is known in contrast to *female;* perhaps something similar was going in this experiment. Early in the mixed block, subjects may have been learning a set of differences between Category B and the previously learned Category A, for example, "has broad wings instead of narrow wings." In this case, the norms for both categories might have been activated for the first few instances of Category B, thus explaining the temporary change in Category A norms.[7]

A third explanation for the temporary increase in listing of Category A defaults following the first instance of Category B notes that the listing of Category A base defaults also increased during the test block, $t(16) = 2.16, SE = 0.038, p < .05$. In Experiment 1, we interpreted a similar rise in listing of base defaults as being due to subjects' using these values to indicate the category membership of each instance during mixed sequences. Perhaps the increase in listing of both Category A defaults and base defaults in the present experiment occurred for this same reason. However, if this explanation was correct, then a similar increase in default listing (and decrease in preference scores) should have been observed in the test block of Experiment 1; no such increase occurred in that experiment. In addition, subjects' listing of base defaults remained elevated throughout the test block of the present experiment, whereas subjects' default listing declined (and preference scores increased) following the first few trials. Thus, it is likely that the apparent decrease in Category A preference scores at this point in training may have reflected some activation of or contrast between the early Category B instances and previous Category A norms. This is an interesting possibility deserving of further study.

Experiment 3

This experiment was a modification of Experiment 2 designed to further investigate category invention in unsupervised learning. In particular, the present experiment investigated the influence of initially aggregating two contrast categories into a single class on subjects' ability to subsequently acquire accurate category-level discriminations.

All conditions of this experiment resembled the contrast condition of Experiment 2, except that the series of same-category instances in the pretraining block was preceded by a single instance from the contrasting category. In the contrast condition of Experiment 2, eight instances of Category A had been presented in succession prior to a mixed block of both Category A and Category B instances. Those eight instances were sufficient for most subjects to learn strong Category A defaults prior to encountering the first instance of Category B,

[7] Such early contrasting of the two categories could not have been detected in Experiment 1 because only instances of Category B were presented during the second block of that experiment. This would have made it impossible to observe any temporary changes in Category A norms during that interval.

thus causing a new category to be created upon seeing the first instance of Category B. In the present experiment, rather than presenting all instances of Category A during the pretraining block, a single instance of Category A was presented on the first trial, followed by a series of instances of Category B (by convention, the category presented first in the training sequence is always referred to as Category A). The main independent variable in this experiment was the number of Category B instances that followed the first instance of Category A in the pretraining block; one group of subjects had 4 instances of Category B in this series, a second group had 8, and a third group had 12. After this pretraining block, a mixed block of both Category A and Category B instances, similar to that of Experiment 2, was presented for the next 13 trials.

The objective of presenting instances from two different categories on the first two trials was to cause subjects to aggregate the categories at the start of training. Because Category A was presented first, the aggregate norms should have initially been dominated by the values of that Category A instance. As subsequent instances of Category B were presented, however, the consistent features of that category should have competed with, and then dominated, the contrasting Category A values in the aggregate norms. If sufficient instances of Category B occurred in this series, these Category B values would be learned as defaults of the combined category, so that presenting a second instance of Category A would trigger a new category to accommodate it. The result of more instances of Category B, then, would be rapid learning of both Categories A and B during the subsequent mixed block.

By contrast, if insufficient Category B instances occurred prior to the test block, the probability of creating a new category should have been reduced. This reduction would result from the relatively high residual strengths of the Category A values in the aggregate norms, which would lessen the perceived contrast between those norms and the features of the first instance of Category A of the test block. If, as predicted, such subjects perceived little disparity and failed to segregate the second instance of Category A from the aggregated norms, that failure would be revealed in their attribute listings during the final mixed block, when they should show reduced learning of both categories.

Autocorrelation models predict a different pattern of results. Consistent with category invention, in such models one would expect that increasing the number of instances of Category B in pretraining should increase later Category B learning, simply because of increased practice. However, in autocorrelation theory one would expect that this manipulation would also decrease later Category A learning because of negative transfer or interference at the level of correlational associations or rules. Thus, the autocorrelation theory is inconsistent with improved learning of Category A because of the increased number of instances of Category B presented during pretraining.

Method

Subjects. The subjects were 36 undergraduate students of Stanford University participating in partial fulfillment of an introductory psychology course requirement.

Procedure. The procedures for this experiment were identical to those of the previous two experiments, except that the numbers of trials differed. Subjects were tested for a single half-hour session in groups of 8 to 10. They were given test booklets similar to those used in Experiments 1 and 2 and were allowed to complete the listing task at their own pace. The listing instructions were identical to those used in Experiments 1 and 2.

Materials and design. The stimuli in this experiment were the same pictorial insect stimuli used in Experiments 1 and 2. The stimulus set was partitioned into categories on the basis of perfectly correlated values on five binary attributes, as in Experiment 2. The remaining three attributes varied independently over two values, different for the two categories. The design shown in Table 2 for Experiment 2 held true for Experiment 3.

The main difference between Experiment 3 and Experiment 2 was the order in which training instances from the two categories were presented. The first instance was always different from the second: following the conventions of previous experiments, we refer to the instance presented first as belonging to Category A. The following n instances were from Category B; the number of instances in this series was the independent variable in this experiment. These first $n + 1$ instances (one Category A instance plus n Category B instances) were referred to as the pretraining block. This pretraining block was followed by a mixed test block of seven Category A and six Category B instances presented in random order (with the constraint that no more than two instances of the same category could occur in a row).

Each of the 16 possible instances from the training set was presented at least once, and instances were selected for a second or third presentation such that each value of the variable attributes appeared equally often. As in Experiment 2, two different stimulus sets were prepared such that assignment of default or variable status to a given attribute was balanced across the group of subjects; this balancing is depicted in Table 2. For both stimulus sets, booklets were constructed such that one category of insects played the role of the first-presented Category A for some subjects, whereas other subjects received booklets in which the other set of insects played the role of Category A. Crossing these two balancing factors (the stimulus set used and the order in which categories were presented) with the three levels of the n variable (number of Category B instances in the pretraining series) yielded a total of 12 groups. Three subjects were randomly assigned to each group, for a total of 36 subjects in this experiment.

Results and Discussion

Preference scores for the three conditions of this experiment are shown in Figure 5. The main prediction of category invention tested in the present experiment was that increasing the number of instances of Category B in pretraining would increase learning of both categories in the following mixed block. This was expected because increasing the number of Category B instances should increase the relative strength of Category B values in the aggregated norms while decreasing the residual strength of Category A values from the first trial. This, in turn, should increase the probability of creating a new category when the next instance of Category A is encountered because these Category A values should appear relatively surprising with respect to these aggregated norms. Once the categories were disaggregated by this triggering, default learning could occur rapidly for each.

The pattern of results shown in Figure 5 lends support to these expectations. Preference scores for Category A (numbered A2 through A8 in Figure 5) increased significantly during the test block for conditions $n = 12$, $t(11) = 3.68$, $SE =$

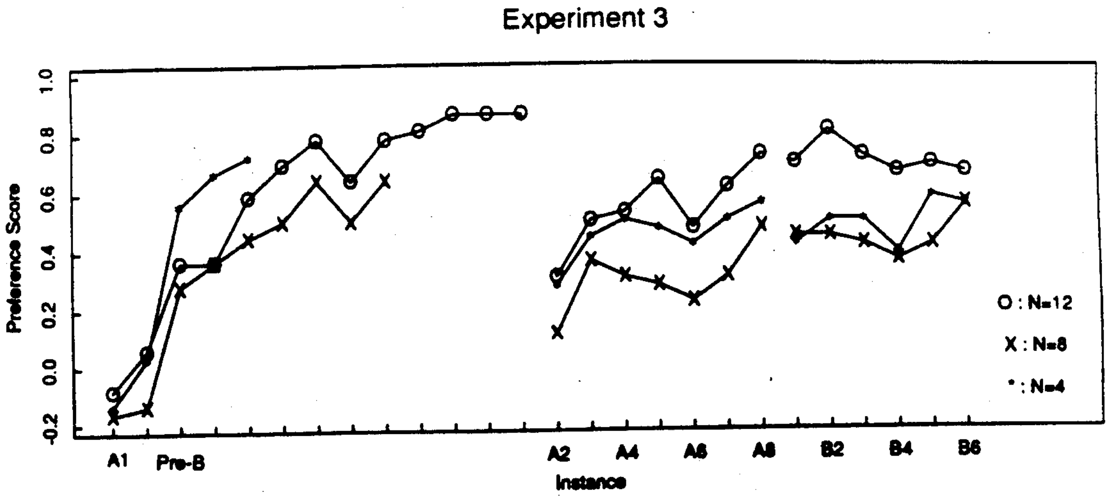


Figure 5. Preference scores for the three conditions of Experiment 3. Pretraining trials are shown in their original order, whereas the test block trials are separated by category.

0.385, $p < .01$, and $n = 8$, $t(11) = 2.48$, $SE = 0.358$, $p < .05$, and marginally significantly for condition $n = 4$, $t(11) = 1.93$, $SE = 0.447$, $p < .10$. This suggests that some learning may have occurred in all three groups. Asymptotic learning of Category A was estimated by averaging preference scores across the last four instances of Category A of the test block. As expected, these averages indicated that learning was significantly higher in condition $n = 12$ than in condition $n = 8$, $t(22) = 3.32$, $SE = 0.088$, $p < .01$. However, the corresponding difference between the $n = 12$ and $n = 4$ conditions failed to attain conventional levels of statistical reliability, $t(22) = 1.22$, $SE = 0.102$, $p > .10$. Learning appeared to be somewhat higher in the $n = 4$ than in the $n = 8$ condition, but this comparison also failed to reach statistical significance, $t(22) = 1.56$, $SE = 0.106$, $p > .10$. When the data from conditions $n = 4$ and $n = 8$ were pooled, the comparison between this combined condition and the $n = 12$ condition was statistically significant, $t(34) = 2.36$, $SE = 0.088$, $p < .05$. Overall, these results indicate higher Category A learning in condition $n = 12$ than in the other two conditions.

Comparisons of Category B learning showed an ordering of conditions similar to those of Category A. Within the pretraining block, learning appeared greater in the $n = 12$ condition than in the $n = 8$ and $n = 4$ conditions, but not greater in the $n = 8$ than the $n = 4$ condition. Preference scores on the last pretraining trial were marginally greater in the $n = 12$ condition than in the $n = 8$ condition, $t(22) = 2.06$, $SE = 0.076$, $p < .10$, nonsignificantly greater in the $n = 12$ condition than in the $n = 4$ condition, $t(22) = 1.47$, $SE = 0.094$, $p > .10$, and not significantly different between the $n = 8$ and $n = 4$ conditions, $t(22) = 0.75$, $SE = 0.111$, $p > .20$. The results appeared slightly stronger when only default listings from the final pretraining trial were compared. Default listing was significantly less in the $n = 12$ than the $n = 8$ condition, $t(22) = 2.27$, $SE = 0.098$, $p < .05$ and in the $n = 4$ condition, $t(22) = 2.24$, $SE = 0.062$, $p < .05$, but there was no significant difference between the $n = 8$ and $n = 4$ conditions, $t(22) = 0.81$, $SE = 0.103$, $p > .25$.

Turning to the test block, one can see that Category B learning was again higher in the $n = 12$ condition and lower in the other two conditions. Preference scores for the $n = 12$ condition exceeded those of the $n = 8$ condition by 27%, a significant difference, $t(22) = 2.14$, $SE = 0.126$, $p < .05$. In addition, preference scores were 22% higher in the $n = 12$ condition than in the $n = 4$ condition, a marginally significant effect, $t(22) = 1.73$, $SE = 0.126$, $p < .10$. No significant difference was obtained between the $n = 4$ and $n = 8$ conditions, $t(22) = 0.41$, $SE = 0.123$. When the $n = 4$ and $n = 8$ conditions were pooled into a single condition, preference scores in this condition were significantly less than those in the $n = 12$ condition, $t(34) = 2.28$, $SE = 0.107$, $p < .05$.

Although some learning may have occurred in all three groups, the stronger learning observed in the $n = 12$ condition favors the category invention theory over a pure autocorrelation theory. The present results, therefore, reinforce and expand on the results of Experiment 2 by demonstrating further patterns of transfer that appear incompatible with strict autocorrelation and that appear to require category invention. However, category invention does not predict higher learning in the $n = 4$ condition than in the $n = 8$ condition, which appeared to have occurred here; rather, we had expected a monotonic increase in learning as n was increased from 4 to 12. The most plausible interpretation of these results is that no real differences existed between the $n = 4$ and $n = 8$ conditions, only between these two conditions and the $n = 12$ condition. Although the $n = 8$ condition appeared to show slightly less learning in some comparisons than the $n = 4$ condition, these comparisons were not statistically significant. Moreover, it appears likely from these data that the baseline learning ability of subjects assigned to the $n = 4$ condition was higher than that of subjects in the other two conditions. When we compared an interval of pretraining trials shared by all three groups (the second- to fourth-presented instances of Category B), we found that learning was significantly higher in the $n = 4$ condition than in the $n = 8$, $t(22) = 3.40$, $SE = 0.084$, $p < .01$ and $n = 12$ conditions, $t(22) = 2.70$, $SE = 0.789$, $p <$

.02, but there was no difference between the $n = 8$ and $n = 12$ conditions, $t(22) = 0.77$, $SE = 0.096$, $p > .25$. This suggests that the unexpectedly high levels of learning observed in the $n = 4$ condition during the test block were likely a spurious outcome of random sampling, which by coincidence assigned better learners to the $n = 4$ condition than to the other two conditions in this experiment.

General Discussion

These three experiments showed powerful effects of the sequencing of training instances on unsupervised learning. These sequence effects are readily interpreted in terms of subjects' inventing partitions or categories for the stimulus domain, but they are not easily accommodated within a simple autocorrelational framework. Both practice (number of instances presented from a given category) and transfer (how instances of one category affect the norms of another) interacted strongly with training sequence in these experiments. Learning of a category normally improved with practice, but only if subjects had explicitly partitioned the instance space. For example, learning of Category B was reduced by presenting more instances of Category B in the pretraining block of Experiment 2, because those extra instances of Category B increased the difficulty of initially separating the two categories. Interference occurred when both types of instances were pooled into a single category because pooling obscured the correlational patterns each contained. However, this interference was eliminated once subjects had learned to separate the categories.

The major procedural difference between supervised and unsupervised learning is that unsupervised subjects must create their own categories and apply these categories without feedback from an external tutor. The category invention problem is thus central to the study of unsupervised learning. Our results provide evidence for a nonincremental, contrast-based category invention process in unsupervised learning. Characterizing details of how this process works, the principles that people use to decide when to create new categories under different training conditions, and how this decision affects other aspects of their performance (e.g., episodic memory for instances) will be a central concern in further studies of unsupervised learning.

A Role for Autocorrelation?

So far, we have argued that the sequence effects observed in our experiments require an explicit process of category invention for their explanation. In particular, the superior learning that occurred when categories were separated in the training sequence was interpreted as being due to this category invention process. However, the residual, or baseline, learning that occurred in mixed training sequences (particularly in Experiment 1), is yet to be explained. If categories must be separated early in training for category invention to occur, then how does the theory explain learning in mixed conditions?

One possibility is that separate categories are created whenever possible but that learning may occur by autocorrelation in other circumstances. Our results do not imply that such autocorrelation is never a factor in unsupervised learning, only that this process alone cannot account for the sequence effects observed here. One possible hybrid theory would incorporate both category invention and autocorrelation; according to such a theory, subjects would normally accumulate some information about interfeature correlations as they processed successive training instances in a discrimination task. This would enable them to learn correlational patterns eventually, even without explicit category invention. Such a correlation learning process might be relatively slow because a large matrix of interfeature correlations would have to be learned in order for a category to be acquired. Consistent with this prediction, learning in the mixed conditions of these experiments was slower than that observed in the blocked conditions.

Although the present results do not eliminate the possibility of explicit autocorrelation in the mixed conditions, the results can also be explained without such autocorrelation. In principle, strict separation of categories in the training sequence should not be required for category invention to occur. For example, it might be assumed that learners in our experiments may invent a new category with some probability, *P*, whenever an instance is presented that is from a different category than the stimulus presented on the trial before (e.g., when an instance of Category B is presented on a trial following one or more instances of Category A). The value of *P* would depend, in part, on the strength of the Category A default values in the norms for the single category that had been applied to all instances up to that point in training. In a blocked sequence, *P* would be high when the first instance of Category B occurred because subjects would have learned strong Category A defaults prior to encountering this instance. In a mixed sequence, *P* would be lower because both Category A and Category B defaults would be encoded as routine values in a set of aggregated norms, and neither would cause a radical mismatch with these norms nor a high probability of inventing a new category when they were presented.

However, category invention could still occur in a mixed sequence so long as the value of *P* was not too low. To illustrate, imagine that there was a 10% chance that the learner would create a new category whenever an instance of Category A occurred after an instance of Category B, or vice versa. Assuming that instances from Categories A and B were presented in alternation, the probability of creating a category on or before the n^{th} alternation is $1 - (1 - P)^n$, which, for $P = .1$, reaches 53% by the sixth alternation. However, category invention would occur at different times for different subjects. Some subjects might discriminate between categories virtually from the start of training, others might do so later in the sequence, and a few might fail to do so by the end of a given training session. The data from such a process, averaged over a group of subjects, would show much the same pattern of apparently gradual learning predicted by the autocorrelation theory.

In summary, these experiments cannot discriminate between pure category invention and a hybrid theory that includes both category invention and autocorrelation. Although the present data provide evidence for the existence of a category invention process, they cannot be interpreted as evidence for the nonexistence of autocorrelation.

Generality of the Results

One objection to generalizing from these results to unsupervised learning in the real world is that the stimulus variation in these experiments was rather artificial and stereotyped compared with the rich, complex variation typical of real-world domains. This objection applies to almost all laboratory research on category learning, which typically uses artificial stimuli generated from combinations of as few as two or three attributes. The purpose of the present experiments was to evaluate the attribute-listing task as an index of unsupervised learning in a relatively simple situation and to use it to make elementary discriminations among models of learning in that situation. Demonstrating that a process such as category invention occurs under artificial conditions constitutes a prefectly valid proof of the existence of that process, although it leaves the issue of boundary conditions unexplored.

In principle, the basic attribute-listing method could be used with many types of stimuli, including stimuli more complex and naturalistic than those used in the present experiments. However, complex stimuli should not change the basic pattern of results (i.e., a shift away from listing predictable aspects of the stimuli with an increasing focus on unpredictable information as categories are learned).

Another sense in which the present stimuli appeared artificial was in the fact that the default values of each category occurred with 100% reliability, that is, attributes were perfectly correlated. These experiments did not attempt to demonstrate unsupervised learning of categories with probabilistic defaults, which may limit the generality of the present results. However, the attribute-listing procedure should be generalizable to learning problems in which category defaults are somewhat unreliable, assuming that people can learn such categories without feedback. It is clear from subjects' performance in the present tasks that many of the fuzzy categories used in standard supervised learning experiments, in which diagnostic features are often highly unreliable (e.g., Estes, Campell, Hatsopoulus, & Hurwitz, 1989; Homa, 1984; Medin & Schaffer, 1978), might be very difficult for subjects to learn without explicit feedback. This difficulty in learning would simply reflect the greater difficulty of the unsupervised learning task itself, which requires subjects to generate their own categories and internal feedback. The investigation of such issues should provide interesting topics for future research and will allow useful comparisons between supervised and unsupervised learning.

References

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 27–90). Hillsdale, NJ: Erlbaum.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451.

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22,* 261–295.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science, 12,* 587–625.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing, 37,* 54–115.

Clapper, J. P., & Bower, G. H. (1991). Learning and applying category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 27, pp. 65–108). San Diego, CA: Academic Press.

Davis, B. R. (1985). An associative hierarchical self-organizing system. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15,* 570–579.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556–571.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 234–257.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery.* Cambridge, MA: MIT Press.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, pp. 49–94). San Diego, CA: Academic Press.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 136–153.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–166). San Diego, CA: Academic Press.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review, 85,* 207–238.

Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 331–364). Palo Alto, CA: Tioga.

Millward, R. B. (1971). Theoretical and experimental approaches to human learning. In J. W. Kling & L. A. Riggs (Eds.), *Experimental psychology* (3rd ed., pp. 905–1017). New York: Holt, Rinehart & Winston.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York: McGraw Hill.

Postman, L. (1971). Transfer, interference, and forgetting. In J. W. Kling, & L. A. Riggs (Eds.), *Experimental psychology* (3rd ed., pp. 1019–1132). New York: Holt, Rinehart & Winston.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

Rosch, E., & Mervis, C. B. (1975). Family resemblence studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 45–77). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Orthony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Hillsdale, NJ: Erlbaum.

Schank, R. C. (1982). *Dynamic memory.* Cambridge, England: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Hillsdale, NJ: Erlbaum.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford England: Basil Blackwell.

Received July 9, 1992
Revision received May 24, 1993
Accepted June 11, 1993 ■

AMERICAN PSYCHOLOGICAL ASSOCIATION
SUBSCRIPTION CLAIMS INFORMATION

Today's Date:_______________

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do NOT duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

____
PRINT FULL NAME OR KEY NAME OF INSTITUTION

____
MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL)

____
ADDRESS

____
DATE YOUR ORDER WAS MAILED (OR PHONED)

____PREPAID ____CHECK ____CHARGE
CHECK/CARD CLEARED DATE:_______________

____
CITY STATE/COUNTRY ZIP

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

____
YOUR NAME AND PHONE NUMBER

ISSUES: ____ MISSING ____ DAMAGED

TITLE	VOLUME OR YEAR	NUMBER OR MONTH

Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4–6 weeks.

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED:_______________ **DATE OF ACTION:**_______________
ACTION TAKEN:_______________ **INV. NO. & DATE:**_______________
STAFF NAME:_______________ **LABEL NO. & DATE:**_______________

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.

Instance and Category Learning in Unsupervised Tasks

John P. Clapper and Gordon H. Bower

Stanford University

ADDRESS ALL CORRESPONDENCE TO:

Gordon H. Bower
Department of Psychology, Bldg. 420
Stanford University
Stanford, CA 94305
(415) 387-5544

Running Head: UNSUPERVISED LEARNING

Abstract

These experiments investigated unsupervised category learning using tasks in which subjects attempted to memorize the features of training instances from two contrasting categories. On each trial, subjects studied a hidden verbal feature list (describing a training instance) one feature at a time for 24 sec, after which they received multiple choice recognition tests to evaluate their memory for each feature. The amount of time spent looking at each feature during the study phase, and the accuracy of recognition during the test phase, provided separate indices of unsupervised learning on each trial. Our main independent variable was the specific presentation sequence (blocked vs. mixed) for instances from the two categories. Blocking produced far faster learning, suggesting that subjects use an explicit "category invention" process, triggered by contrasting examples, to capture the correlational structure of the stimulus domain. The results also demonstrated the selective encoding and enhanced memory for instances predicted by schema-based theories of learning. These results held true for categories defined by probabilistic as well as deterministic default features.

Instance and Category Learning in Unsupervised Tasks

The ability to learn and use categories is fundamental to human intelligence. Categories inferred from examples may be acquired under two general types of training conditions, referred to as *supervised* and *unsupervised* learning. In a typical supervised learning experiment categories are defined in advance by an experimenter who also provides relevant feedback (correct answers) so that subjects can gradually learn to match their categories to the correct class of training instances. By contrast, in unsupervised learning tasks subjects receive neither predefined categories nor feedback from an external tutor. Rather, subjects must discover categories for themselves as they examine a series of training instances, basing such categories on any patterns or regularities they observe among these stimuli.

An extensive research tradition has accumulated in the study of supervised learning (see, e.g., Bruner, Goodnow & Austin, 1956; Millward, 1971; Smith & Medin, 1981), but there have been comparatively few empirical studies of unsupervised learning. One reason for this paucity of research may have been a lack of reliable measures of category learning within such tasks. For example, the primary measure used in studies of supervised learning - accuracy in assigning an instance to one of the predefined categories - is by definition inapplicable to unsupervised learning.

Clapper and Bower (1991, 1994) developed and tested an index of unsupervised learning which they called "attribute listing". The present article introduces a second method for investigating unsupervised learning. In addition to providing information about the abstraction of category norms from a series of training instances, the method also provides information about how subjects' discovery of a category alters and economizes the way in which they process and remember individual instances.

Defining Categories in Unsupervised Tasks

Unsupervised learning can be defined in terms of subjects' ability to detect and learn about pre-existing structure or patterns within a set of training stimuli. Therefore, to investigate such learning we should first describe the kinds of "patterns" or "structure" we consider as giving rise to distinct categories, thus enabling us to evaluate subjects' learning of them.

In this article, categories will be defined in terms of correlated (consistently co-occurring) properties among training instances within a stimulus set. We adopt the conventional vocabulary for describing training instances in terms of abstract dimensions or *attributes*, each of which can assume a number of concrete *values* (Clapper & Bower, 1991, 1994). For example, people differ in the attribute of hair color, with blond, brown, red, and black being possible values of this hair-color attribute. The specific value of an attribute in a given instance is also referred to as a *feature* of that instance. In principle, attributes may be either additive (with two values, present or absent) or substitutive (with any number of alternative values, such as different hair colors; see, e.g., Tversky, 1977). Attributes may also be discrete or continuous (e.g., ordered dimensions such as height or weight). In this article, only the discrete, substitutive case will be considered, although the methods described should also be applicable to other cases.

For such stimulus domains, categories of instances may be defined in terms of correlations among the values of these attributes (see Figure 1). Such correlational structure provides an inductive basis for partitioning a domain into separate categories, each corresponding to a particular set of correlated features (Garner, 1974). For example, in Figure 1a, Category A is characterized by correlated values (denoted by a 1) on the first five listed attributes, while Category B has different correlated values (denoted by a 2) on the same attributes. Category members differ in terms of the last three uncorrelated attributes listed and these vary freely within each category.

Insert Figure 1 about here

Importantly, such correlational structure provides a learner with predictive power: given that one correlated value is observed, the presence of the other four correlated values can be inferred. This predictable structure may be contrasted with the stimulus set illustrated in Figure 1b, in which all attributes of the stimuli vary independently so there is no natural, informative way to partition the set into distinct categories. In this uncorrelated set, knowing some attribute values of an instance does not improve the learner's ability to reliably predict any of its other values.

A second example of categories defined in terms of correlational structure is shown in Figure 1c. Here, the values of all the attributes are correlated, but unlike Figure 1a, the attributes in Figure 1c are only imperfectly correlated. The correlational patterns define two distinct categories, but the characteristic features of those categories occur probabilistically, rather than deterministically, as in Figure 1a. However, even less than perfect correlations can provide some predictive power, and many natural categories appear to have probabilistic, rather than deterministic, features (Wittgenstein, 1953; Rosch, 1975, 1977).

The correlated attribute values that characterize a given category will be referred to in this article as its *default* features. The term "default" will be used whether the correlated features occur with perfect reliability, or just with high probability. An uncorrelated attribute that varies independently of the other attributes in its category will be referred to as a *variable* attribute, and its specific values will be referred to as variable values. An unusual or improbable value occurring in place of an expected default value (as in Figure 1c) is dubbed an *exception* or *default violation*, since it violates the correlational regularities that characterize the collection of instances.

Measuring Unsupervised Learning

Defining categories in terms of correlated values, a procedure can then be designed to index subjects' learning of a given category as they experience successive training instances. In considering such tasks, we may distinguish between *direct* measures of categorization, in which subjects are explicitly instructed to group (or sort) the training instances into salient categories and learning is evaluated in terms of subjects' success at this task, and *indirect* measures in which categorization is not an explicit goal but category learning is evaluated indirectly by its influence on how subjects perceive, evaluate, or remember individual training instances. An example of the former task would be experiments in which subjects sort a set of stimulus cards into some "natural" groupings (e.g., Miller, 1969). The latter approach will be illustrated by the experiments described below, as well the "attribute listing" procedure described in Clapper and Bower (1991, 1994).

We studied the learning of correlation-based categories using tasks in which subjects' goal was simply to *memorize* individual training instances; in fact, category learning was never mentioned to subjects until their debriefing at the end of the experiment. Subjects saw training instances composed of many attributes; some attributes had correlated values (defining two contrasting categories), and others did not. Instances corresponded to lists of several verbally described features, supposedly belonging to different (fictitious) trees, presented to subjects on a computer screen. For example, a given tree might be described as having dark grey bark, a high commercial value, fast growth, and so on. Subjects were required to study each of these instance descriptions (or feature lists) for a fixed study-time in order to prepare for a memory test about it. During this time, the computerized display was arranged so that the subject could look at only one feature at a time, although the subject could choose to inspect different features for a self-selected time. After the study period, subjects received a multiple choice test of their ability to recognize which features had occurred in the previous instance. Two types of data were collected: (1) the time devoted to examining each attribute value during the study period, and (2) the accuracy of remembering each value of the instance during the testing phase that immediately followed its study.

If subjects in this task learned the categories (correlational patterns), the default features of the individual training instances would become predictable. As a result of this increased predictive power, subjects' overall memory for the instances should increase. Once subjects learned the patterns of correlated defaults, they should be able to remember all the default features of an instance merely by remembering its category membership (or a single default value diagnostic of that membership). Even if subjects were unable to retrieve a given default value from their memory of the most recent instance, they could infer its probable presence based on generic category norms abstracted from previous instances.

In addition to this improvement in guessing defaults at the time of test, memory for variable attributes should also improve due to changes in subjects' attentional priorities during the study period. As subjects learn the defaults, they require less time to encode them. Consequently, a rational learner with a limited encoding capacity and limited time should assign a higher priority to learning variable or exceptional values of instances, since these features of the instance cannot be inferred from category norms. Thus, as more instances are seen, subjects learning the correlations of defaults within categories should also show progressively more study and better memory for the values of variable attributes, or for exceptional values that occur in place of an expected default, compared to control subjects for whom all attributes of the stimuli are uncorrelated. In essence, subjects were expected to adopt an "uncertainty reducing" strategy when learning the training instances, assigning highest priority to encoding those features of an instance that are least predictable from category norms. These effects on attentional allocation were measured directly in the present task, since the computer recorded how long subjects spent studying each feature of the instances.

This task has several attractive properties as a method of studying unsupervised category learning. First, facilitating episodic learning about specific objects or situations is one important function of category knowledge. This adaptive function is illustrated by many experiments which show that people better remember textual or pictorial information when they can relate it to schematic knowledge about familiar scripts or categories (e.g., Bransford & Franks, 1971). Similarly, experts in domains such as electronics or chess are able to use their knowledge of general patterns (categories) within those domains to remember stimuli or situations (e.g., a particular electronic circuit or chess board configuration) much more accurately than novices lacking such categories (e.g., Chase & Simon, 1973; deGroot, 1965, 1966). Given the sharply limited capacity of people to learn random or meaningless material (e.g., lists of unconnected facts), and the complexity of many naturalistic learning problems, this facilitation must be extremely important to people's ability to function intelligently.

A second advantage of studying category acquisition in the context of learning individual instances is that such tasks make it possible to investigate *incidental* category learning, in contrast to the intentional hypothesizing strategies typically adopted by subjects in direct tasks which involve overtly sorting or classifying the patterns. The present task enables the investigation of categories or patterns that "pop out" of the learners' stream of experience, as opposed to those they are able to uncover only by deliberate search. A significant portion of people's commonsense knowledge of the world is probably acquired in such incidental fashion, although we know of no empirical studies that directly address this issue.

Third, the effect of category knowledge, in the form of schemas, scripts, or stereotypes, upon episodic memory is an important research topic in itself (e.g., Bower, Black, & Turner, 1979; Graesser, Woll, Kowalski, & Smith, 1980; Srull & Wyer, 1989). Most of this research has employed naturally acquired categories (e.g., the familiar "restaurant script" of Schank & Abelson, 1977), and investigated memory for text or pictures based on these categories. The present task offers a method of studying these issues using artificial category knowledge synthesized in the laboratory under controlled conditions. This greater control may eventually enable the investigation of factors that would be difficult to study using only naturalistic materials.

Theoretical approaches

While many techniques of conceptual clustering have been proposed in the literatures of numerical taxonomy, psychological scaling, and artificial intelligence, most of them have not been formulated as incremental learning procedures (Everitt, 1980; Anderberg, 1973). We will postpone consideration of these clustering ideas until the final discussion by which time we will have more facts in hand to aid in their evaluation.

For immediate consideration, we will describe two general classes of incremental theories regarding how unsupervised categories might be learned. One of these is based on the idea of cumulative learning of inter-feature correlations; the other is based on the idea of discrete invention of subjective categories about which norms are acquired.

Autoassociation models. One class of models that are applicable to the unsupervised category learning task are autoassociators, often formulated within a connectionist architecture. We consider these as serious, incremental models of human category learning, and will examine how they might apply to our situation. We will contrast these autoassociation models to another class which we will champion, dubbed "category invention" models.

As their name implies, the autoassociators assume that subjects directly accumulate evidence about pairwise feature correlations as successive instances of a category are encountered. This approach is illustrated by the one-layer connectionist autoassociation models of J. A. Anderson (Anderson, 1977; Anderson, Siverstein, Ritz & Jones, 1977) and McClelland and Rumelhart (1985; Rumelhart, Hinton, & McClelland, 1986). It is also instantiated in rule-based systems such as those of Billman and Heit (1988) and Davis (1985). By keeping a record over instances of the co-occurrence frequencies of all (or many) possible pairs of attribute values, a learner could capture the pairwise correlational structure of stimulus sets such as those in Figure 1 without necessarily partitioning the domains into explicit categories. Information typically provided by such a categorization would be implicit in an exhaustive co-occurrence record; in fact, explicit categorization might actually lose or obscure certain co-occurrence information if individual co-occurrences were to be replaced by the average likelihood that each attribute value occurs within the category.

A category invention hypothesis. A second method of learning new categories in unsupervised tasks postulates a discrete "category invention" process. In this method, subjects acquire co-occurrence patterns by first hypothesizing a partition of the stimulus set into separate, explicit categories corresponding to these patterns. Descriptive norms or expectations about feature probabilities within each category are then stored in separate data structures, such as prototypes or schemas (e.g., Posner & Keele, 1968; Reed, 1972; Minsky, 1975; Rumelhart & Ortony, 1977; Schank & Abelson, 1977; Schank, 1982; J.R. Anderson, 1991). By sorting stimuli containing different co-occurrence patterns into different categories, and then computing conditional frequency distributions within these categories, a learner could capture much of the same information contained in a direct correlational record.

The major issue for the category invention approach is deciding when, and on what basis, new categories should be invented. All theorists agree that learners are practically forced to rely on the match or mismatch of each stimulus to existing categories to decide whether or not to invent a new category (e.g., Schank, 1982; Holland, Holyoak, Nisbett, & Thagard, 1986; Lebowitz, 1987; Fisher, 1987; J. R. Anderson, 1991). Assuming that subjects create a new category at the start of the experiment to describe the first training instance, they should continue assimilating further instances to this category until an instance is encountered that contrasts (mismatchs) sufficiently with previous category norms to justify the creation of a distinct second category. Any later instances similar to this initial "triggering" instance would then also be assigned to the new category and should not affect subjects' norms about the first category. Separating expectations about different categories in this way would allow new patterns to be learned without discarding or distorting norms learned about previous categories.

Comparing Autoassociation to Category Invention

Since learning by category invention depends on contrast while learning by autocorrelation depends on practice (i.e., accumulating evidence about feature correlations over multiple instances), it is not surprising that they expect different outcomes from certain types of experimental manipulations. In the present experiments, we focus primarily on manipulations of the particular *sequence* in which instances of different categories are presented. Specifically, we argue that a process of category-invention-by-contrast should be more sensitive to manipulations of training sequence than would be an autocorrelation learner. Moreover, the particular pattern of sequence effects predicted by category invention cannot be easily reproduced by models restricted to autocorrelation. If obtained in our experiments, the predicted sequence effects would constitute strong evidence that our subjects were using stimulus contrasts to invent new categories. These experiments, if successful, would constitute an "existence proof" of a discrete, non-incremental category invention process in unsupervised learning. (However, the results could not be interpreted as strong evidence for the *non*-existence of autocorrelation. We will return to this issue in the General Discussion).

Clapper and Bower (1994) reported several experiments which varied the sequence of instances from two different categories, and the results implied that subjects were using perceived contrast to invent separate categories. In one experiment, instances of contrasting categories such as those shown in Figure 1a were presented to subjects whose task was to list a few distinguishing features of each instance ostensibly in order to prepare for a later memory test over the specific instances. One group of subjects in the first part of the experiment saw instances blocked by category, i.e., 12 instances from one category (referred to as "Category A") were presented prior to 12 instances from a second category ("Category B"). Following these two single-category blocks, a mixed block containing an equal number of test instances from both categories was presented in random order. A second group of subjects saw the same final test block, but the 24 instances from the two previous blocks were randomly intermixed rather than separated in the training sequence as in the Blocked group.

A category invention process predicts that the probability of creating a separate category for the first instance of Category B should depend on its perceived mismatch with the norms derived from prior Category A instances. For example, after seeing only one or two instances of Category A, subjects' expectations or norms about this category would probably be rather vague and general; not enough data has yet been seen to begin separating default from variable attributes. If an instance of Category B were presented at this time, subjects might not perceive a *confident mismatch* between this instance and the norms for the previous instances of Category A. In the absence of a confident mismatch of the current pattern with previous norms, subjects are likely to interpret the changed defaults as simply variable attributes; they would thus assimilate the first instance of Category B together with prior instances of Category A, resulting in a single, overgeneralized category that would fail to capture the conditionalized covariations of the default attributes. Further instances from either the A or B categories would then appear consistent with this overgeneralized category. Due to this lack of contrast, subjects should have difficulty abandoning this aggregated category and correctly partitioning the stimulus set into separate categories corresponding to the conditionalized patterns of co-occurring features.

In summary, a category invention process, which triggers new categories only as old ones fail badly, expects that subjects should have difficulty learning with mixed training sequences. The same categories should be distinguished more easily when instances are presented in a blocked training sequence, because strong default norms could be learned for the first category before the second was encountered, thus highlighting the contrast between them. On the other hand, autocorrelation models do not necessarily expect superior learning when instances from different categories are separated in the training sequence, compared to being presented in mixed alternation. Within the autocorrelational approach, mismatch or perceived contrast does not directly affect learning because learners are assumed simply to increment inter-feature correlational strengths based on the features of each instance. Instead, learning of correlations should depend primarily on practice, i.e., the number of times particular pairs of features co-occur across different instances. If learning is determined primarily by the number of instances presented from a given category, then the particular sequencing of instances (e.g., blocked vs. mixed) should be relatively unimportant to such a learning process. Indeed, if the two categories of patterns are orthogonal, then learning will be the same regardless of the sequencing of the sets of instances.

If poor learning is found in the mixed condition, an autocorrelation model might explain it by assuming substantial interference or negative transfer between Category A and Category B. This would arise, for example, if the network coding of the two categories of instances overlapped to some degree. Given such interference, a category might be learned more slowly in a mixed sequence with another category than if it were presented alone. However, such an autocorrelation-with-interference hypothesis would also expect the second-learned category to interfere greatly with the first-learned in a blocked sequence, similar to that observed when subjects learn separate (blocked) lists of paired associates in verbal learning experiments (e.g., the A-B, C-D paradigm of Postman, 1971; Millward, 1971). Thus, learning the correlations of the second category should cause retroactive interference and reduce subject's memory for the correlations of the first category. In connectionist autoassociators, this retroactive interference may be particularly severe, a phenomenon sometimes referred to as "catastrophic interference" (McCloskey & Cohen, 1989; Ratcliff, 1990). For the same reasons, learning the correlational associations of the first category should result in negative transfer and slow learning of the correlational associations that comprise the second category.

Clapper and Bower (1994) found that categories were learned appreciably faster in blocked than in mixed training sequences, and found no evidence for substantial proactive or retroactive interference in blocked sequences. Those results suggested that the slower learning in a mixed-sequence condition was due to a lack of contrast rather than to interference in correlation learning. The result also provided evidence for a discrete category invention process. One purpose of the present experiments was to obtain converging evidence for the category invention hypothesis using a different measure of instance processing.

Experiment 1

The first experiment employed a similar contrast between blocked and mixed training sequences to that described in Clapper and Bower (1994) and it had two goals. A first goal was to provide evidence for the basic validity of the instance memory procedure, described above, as a method of investigating unsupervised category learning. If the results of this experiment agreed with those of Clapper and Bower (1994), we could be more confident of the validity of both tasks and the reality of the underlying processes they presume to assess. The generality of our methods and theoretical conclusions would be bolstered by the fact that the present experiments differed from the earlier studies both in the nature of the task itself (instance memory vs. attribute listing) and in the type of stimuli. To add generality, the present studies used verbal stimuli whereas the earlier ones employed pictorial stimuli. In addition, the stimuli in the present experiment contained a larger number of attributes than did those in previous experiments (12 instead of 8). It is important to compare verbal and pictorial stimuli in research on unsupervised learning because previous research indicates that verbal stimuli may be remembered and compared differently than pictorial stimuli, which could also mean that they might be categorized with somewhat different strategies (Paivio, 1971; Kosslyn & Pomerantz, 1977; Gati and Tversky, 1984).

Our second goal was to provide further evidence to discriminate the autocorrelation vs. category invention hypotheses. The earlier attribute listing studies provided evidence for category invention, and we hoped to replicate this support in the present experiment.

Method

Subjects

The subjects were 43 undergraduate students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement.

Materials

The training instances were verbal descriptions of 32 fictitious trees, presented in a column-list format. The instances were characterized in terms of twelve substitutive attributes, each with four possible values, defining a stimulus set of 4^{12} distinct instances. Examples of these attributes included the color of the tree's bark (dark grey, deep brown, mossy green, or light tan), its form or overall shape (low and shrub-like, tall and column-like, massive and wide branching, twisting and vine-like), the season in which it flowered (spring, summer, winter, or autumn), and so on. For nine of these twelve attributes, only two of the four possible values were presented in the training instances, although all four values appeared as alternatives in the multiple choice tests.

Procedure

Subjects were tested in groups of 10 to 15 in a computer lab for a one-hour session. Each subject was seated at an individual microcomputer terminal that administered the entire experiment. After subjects read the instructions presented on the computer screen and signed a form indicating their informed consent to participate, the main portion of the experiment began.

Each trial consisted of a study phase followed by a test phase. At the beginning of the study phase, a verbal list was presented in the middle of the CRT screen. At the top of the list was the name of a fictitious tree instance (these were arbitrarily selected Latin names from a plant identification guide), below which appeared a list of twelve verbal feature descriptors, one in each of 12 rows. (Every tree-instance had a different name, so there was no suggestion that the name referred to a class or family of species). Each row contained a verbal description of a specific value of a particular stimulus attribute. The attributes were presented in the same serial order (screen locations) on each trial, different for each subject.

At the start of the trial, each descriptor was masked by a row of X's (see Figure 2a). Starting with the cursor at a random position in the list, subjects could look at the descriptors by pressing a designated "line down" key or a "line up" key which removed all the X's and allowed subjects to examine the item (attribute value) on that one line. The exposed attribute value on a given line was covered up again as soon as the subject moved the inspection pointer to a new line (attribute). This procedure permitted subjects to examine the features in any order they wished, and to spend as much time as they wished studying any particular feature, although they were limited to a total study-time of 24 sec for the entire 12-item list. The computer recorded the total amount of time a subject spent looking at each attribute of each instance.

Insert Figure 2 about here

In the test phase subjects were tested for their memory for the values of all twelve attributes of the preceding instance. The twelve test questions were presented one at a time in random order in a multiple-choice format (see Figure 2b). The name of the most recent tree-instance appeared at the top of the multiple-choice display, with four alternative answers below. These alternatives were the 4 different values of an attribute. Subjects tried to remember which of these values occurred in the last-studied instance and typed in the number corresponding to that choice on their computer keyboard. Following this response, the computer displayed either a "correct" or an "incorrect" prompt under the test display, which remained on the screen. If the response was incorrect, the correct choice was indicated by an arrow in the display (see Figure 2c). The subject then pressed the "Return" key to see the next test question.

After answering the twelve test questions about a given instance, subjects received summary feedback for the trial. The percentage of items answered correctly on that trial was displayed, and below this the average percentage correct pooled over all test trials completed up to that point. If the trial score was higher than the cumulative score, the message "Good job! You beat your overall score!" appeared on the screen; if not, the message "Try to beat your overall score next trial" was displayed. If the subject answered all the test questions correctly on a given trial, the message "Good job! Your score was perfect!" was displayed. By pressing the Return key, subjects moved on to studying the next instance in the training series.

The twelve attributes were tested in a different random order on each trial, and the order in which values were listed in the multiple-choice display was also randomized separately on each trial. The experiment consisted of a total of 32 such instance study-test trials. Following this, subjects read a debriefing sheet that informed them of the purpose and methods of the experiment.

Design

Subjects were randomly assigned to three different conditions. In the two correlated conditions the values of the nine two-valued attributes were perfectly correlated across different training instances. The instances could be partitioned into two distinct categories based on these correlated values. Using the notation of Figure 1, the categories were Category A = 111111111xxx and Category B = 222222222xxx, where the first nine attribute positions denote the correlated defaults, and the x's indicate uncorrelated attributes that vary independently through all four values across different instances within a category.

The two correlated conditions differed in the order in which the first 24 of 32 instances were presented. In the *Blocked* condition, the first twelve instances were all members of Category A and the second twelve instances were members of Category B. The final eight trials provided a test series consisting of four A-instances and four B-instances presented in a randomly intermixed sequence. The *Mixed* condition differed from the Blocked condition only in the order in which the first twenty-four instances were presented, i.e., in a randomized sequence rather than Blocked by category. The randomization procedure was so constrained so that no more than three instances from the same category appeared in a row. The final eight trials for the Mixed condition were identical to those of the Blocked condition.

The uncorrelated or *Control* subjects saw instances without categories because all the attributes of the training instances varied independently. As in the correlated groups, nine of the twelve attributes varied through only two values in the training instances, while the remaining three attributes varied through four values. But, because the attribute values in this condition were uncorrelated, there was no structural basis for partitioning the stimuli into separate categories.

The final eight (test) instances presented in the Control condition were identical to those of the two correlated conditions, viz., had correlated values. This final block of correlated instances will be referred to as the *test block* for all three groups. Subjects in the Control condition were expected to show hardly any learning during this test block, providing a baseline against which to evaluate any learning observed in the other two groups.

Counterbalancing

The stimuli for all subjects in a given condition were generated by the testing program from the same input file, which contained coded specifications for generating the instances presented on each trial. Stimuli generated from these codes were presented in the same file order for all subjects in all three conditions, allowing unbiased comparisons of learning across different groups. The correspondence between serial positions in the codes and the order in which an attribute was listed on the computer screen in the training instances was randomized for each subject. These random assignments were undertaken to balance out any idiosyncratic effects of particular attributes, values, or combinations of values, as well as serial position in which the attributes were presented.

Results and Discussion

The two dependent variables recorded on each trial of this experiment were (1) study-times for default and variable attributes during the study phase, and (2) recognition-memory accuracy for defaults and variables during the test phase. The data for this experiment is shown in Figure 3. The main indicators of category learning are (1) subjects' overall accuracy of remembering the attribute values of each instance, computed by averaging the accuracy of remembering defaults and variable features; (2) subjects increasing the time they spend studying variable rather than default features of each instance as categories are learned; and (3) an increase in memory for variable features, probably caused by this increased study-times during the encoding phase.

Insert Figure 3 about here

The category invention hypothesis predicts that category learning should be greater in the Blocked than in the Mixed condition, and, of course, expects no category learning to occur in the uncorrelated, Control condition. These expectations were largely confirmed by the present data. Data analysis in this and later experiments involves large numbers of comparative t-tests of significance. Rather than recite a flurry of t-statistics, as a favor to readers we will adopt throughout a $p < .01$ criterion for significance and simply state which comparisons are or are not statistically significant by that criterion. Readers interested in actual t's and dfs may consult the authors.

Memory Accuracy

Figure 3a for the Blocked condition indicates that recognition memory for instances' values improved rapidly over the first several instances of both Category A and Category B. Averaged across defaults and variables, overall accuracy increased significantly from 0.66 on the first instance of Category A to an average of 0.93 on the last two trials. Accuracy dropped significantly to 0.71 when the first instance of Category B was presented on the 13th trial. But a similar pattern of increasing accuracy was then observed over succeeding instances of Category B, with accuracy increasing and leveling off around 0.97 (averaged over the last 6 instances of the Category B Block).

The Blocked subjects showed a slight drop upon encountering the first A instance in the mixed test block, and overall memory performance during this block was somewhat lower (reliably) than performance during the preceding Category B-block (comparing the average of the last 6 trials of the B-block to the average of the 8 test trials). However, when the first A-instance was excluded from the test scores, the memory performance for A and B instances was about the same during the test trials. Thus, there was only slight evidence (one trial) for any retroactive interference of learning Category B upon memory for the defaults of Category A.

In contrast to the category learning in the Blocked condition, the instance memory data showed little evidence of learning in either the Mixed or Control conditions; in fact, the latter two conditions did not differ reliably. Memory accuracy in the Blocked condition was significantly greater than in both the Control and Mixed conditions. These comparisons remained significant when restricted to the final 8 test trials; averaged over the test block, accuracy percentages were 0.24 higher in the Blocked than in the Mixed condition, and 0.29 higher than in the Control condition.

Study Time Results

The study-time data revealed much the same pattern of significant differences as the recognition accuracy data. In the Blocked condition, the mean study-time pooled over all 32 trials was 1.78 seconds for defaults significantly below the 2.91 sec observed for variable attributes. Examining the mean difference scores plotted over trials in Figure 3b, the preference for studying variable attributes clearly increased throughout the Category A block, from .18 sec on the first trial to 2.01 sec on the twelfth and final trial of this block. Moreover, the preference for studying variables was still increasing at the end of this first block.

Upon shifting to the B-block, the preference index dropped significantly on the first B-trial, from 2.01 on the final trial of the A-block to -0.125 sec on the first B-trial. This coincided with subjects' increased inspection of the novel values of the default attributes, taking time away from inspecting variable attributes. Learning seemed to occur somewhat more rapidly during the Category B-block, so that preferential inspection times had leveled off after the sixth B-instance. Comparing the a and b panels of Figure 3 for Category B, memory accuracy appears to have leveled off on about the same trial as did differential study-time to defaults versus variables.

The Blocked subjects' preference for studying variables decreased somewhat when the first A-instance was presented during the mixed test block, compared to the average of the preceding six B-instances. However, the results in Figure 3b clearly show that the preference for studying variables over defaults remained high throughout the test block. This result is important because it indicates that the apparent learning observed earlier in the blocked training sequence was not merely due to localized habituation to "consecutive" default values, but rather to subjects acquiring and retaining stable norms for the two categories.

Consistent with the recognition memory data, the study-time data for Blocked subjects showed no evidence of retroactive interference due to Category B trials upon performance with Category A. Excluding the surprising first A-instance of the test block, instances of the two categories yielded about the same study-time preference scores throughout the test block. The slightly lower preference scores for instances of both categories during this test block, compared to the six preceding B trials *($p < .02$)*, probably reflect subjects' need to sample enough of the default features to confidently categorize the instances of the test block. In contrast, during the earlier blocks, when category membership was constant over blocks of trials, subjects could spend less time checking the default features of each instance.

Turning to the Mixed condition, variable and default study-times were nearly equal (means of 2.04 and 2.07 sec, respectively) as reflected in the preference scores (in the lower panel of Figure 3b) hovering around zero with no apparent trends. The data for the uncorrelated Control condition were similar to those of the Mixed condition in showing no learning trends. Study-times for variable features averaged only about .06 sec greater than default study-times, a non-significant difference.

Between-group Comparisons

Consistent with these within-group analyses, direct comparisons between groups provided further evidence for superior learning in the Blocked condition. The average study-time preference score of 1.14 sec observed in the Blocked condition was significantly greater than the 0.06 sec and 0.03 sec effects observed in the Control and Mixed conditions, respectively. Subjects in the Blocked condition, in contrast to the other two groups, significantly increased their study-time to variable attributes as they acquired default norms for the two categories. Correspondingly, their memory for variables averaged 0.23 better than that of subjects in the Mixed condition, and 0.24 better than that of subjects in the Control condition.

Subjects showed higher memory for defaults than for variables in all three conditions. Within the Blocked condition, memory for defaults (0.93) was reliably higher than that for variables (0.83). In the Mixed condition, defaults were remembered with an average accuracy of 0.65, which is reliably higher than the variables' accuracy of 0.60. The Control group also showed significantly higher memory for attributes with 2 values (the "default equivalent") than those with 4 values (0.65 vs. 0.58).

This greater memory for defaults in the Mixed and Control conditions could have been due to: (1) Mixed subjects' ability to retrieve correlated default values from their category norms, whereas the values of variable attributes had to be recorded afresh for each instance, or (2) higher guessing of the correct value of two-valued "default" attributes presented during the study phase, compared to the four-valued "variable" attributes. Since there was no other evidence of default learning in the data for the Mixed and Control conditions, the differences obtained in these two conditions were probably due to the guessing factor, and their order reverses when choice percentages are corrected for guessing.

The increased memory for both defaults and variables in the Blocked condition indicates that category learning facilitates encoding of both predictable and unpredictable features of instances. This result replicates earlier ones showing that category knowledge improves memory for both default and non-default properties of instances (Clapper & Bower, 1991), and supports the encoding assumptions of schema theories. Such theories usually assume that learners focus on those aspects of an instance that are surprising or unpredictable with respect to norms stored in the category schema, while backgrounding expected defaults (see, e.g., Bower, et al., 1979; Graesser, et al., 1980). This pattern was observed in the study-time data from the present experiment, and the recognition memory data provided further verification.

To summarize, the pattern of results from both study-times and recognition-memory accuracies support the category invention approach in that learning in the Blocked condition was much better than in the other two groups. There was little evidence for negative transfer due to learning Category A upon subsequent learning of Category B in the Blocked condition; in fact, learning of the second category was achieved as quickly as the first. Nor did Category B learning in the Blocked condition cause more than minor interference in remembering Category A. The autocorrelation-plus-interference hypothesis expects that Blocked subjects' performance on Category A instances during the test block should have been greatly reduced by retroactive interference from interpolated learning of Category B. However, after the momentary surprise of seeing the first A-instance in the test block, subjects performed equally well with the two categories during the remaining tests. This absence of interference contradicts a prediction of autocorrelation theories, i.e., if interference occurs between categories in a mixed sequence, then it should also occur for learning in a blocked sequence. These results are difficult to explain within a strictly autocorrelational theory, and imply that people in unsupervised learning tasks accommodate stimuli that mismatch existing category norms by inventing new categories.

Experiment 2

Experiment 1 demonstrated the utility of the instance-memory task as a measure of unsupervised learning for categories distinguished by deterministic defaults, i.e., perfectly correlated attribute values. Experiment 1 also provided evidence that subjects used a discrete category invention process based on perceived contrast to learn these correlational patterns. These results are consistent with earlier results of Clapper and Bower (1994), in which similar evidence for discrete category invention in domains with deterministic correlational patterns was obtained using the attribute listing task. Experiment 2 attempts to extend these results to categories characterized by *probabilistic,* rather than deterministic, defaults.

This experiment was similar to Experiment 1 comparing Blocked vs. Mixed vs. Uncorrelated Control conditions, except the categories were defined by probabilistic feature correlations similar to those shown in Figure 1c. All the attributes of the stimuli in Experiment 2 were correlated and there were no consistently variable attributes. Thus, the features of an instance were either defaults or exceptional values (default violations). Instances within a category differed in the number of exceptional values they had (0, 1, or 2), and the particular attributes which had exceptional values. As in Experiment 1, learning in the different conditions could be ordered in terms of subjects' ability to remember both default and non-default features of the instances, as well as their tendency to study non-defaults longer than defaults during the study phase of each trial.

We expected the results of this experiment to be similar to those of Experiment 1. First, subjects were expected to demonstrate unsupervised category learning in the correlated groups (particularly in the Blocked condition) compared to the uncorrelated control group. Such a result would demonstrate the possibility of unsupervised learning of categories with probabilistic defaults, thus enhancing the utility and generality of the current instance-memory task and our general approach. Second, subjects in the Blocked condition were expected to show better learning than subjects in the Mixed condition, since blocking highlights the contrast between the two correlational patterns, thus facilitating the invention of separate categories to describe them. Such a result would provide further evidence for the importance of discrete category invention in unsupervised learning.

Method

Subjects, Materials, and Procedure

The subjects were 36 students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement. As in Experiment 1, the training instances were verbal descriptions of fictitious trees presented in a list format on a computer screen. Each instance was described by 12 attributes, each of which had 4 possible values. The experimental procedure was identical in most respects to that of Experiment 1 with study-time and recognition memory data collected on each instance. The experimental session consisted of 36 trials plus instructions and debriefing. Subjects were tested in groups of 10 to 15 for a single session lasting approximately one hour.

Design

Subjects were randomly assigned to 3 different conditions, similar to those of Experiment 1. In 2 of these conditions, the values of all 12 attributes were strongly (but not perfectly) correlated, such that the stimulus set could be partitioned into 2 distinct subsets or categories based on these correlated values. These categories may be denoted as Category A = 111111111111 and Category B = 222222222222.

Within each category, one-quarter of the instances had no exceptional values, one-half had a single exceptional value, and the one-quarter had 2 exceptional values. Since each instance had 12 attributes, these variations imply that default values occurred with an average probability of 0.92 within a given category. Exceptional values occurred equally often on all 12 attributes.

Different exceptional values were used for the two categories as illustrated in Figure 1c. To illustrate, instances of Category A, in which value #1 was the default, had value #3 as the exceptional value (e.g., 111311111111), whereas in Category B, value #2 was the default and value #4 was exceptional, (e.g., 4222224222222). These numerical codes were chosen arbitrarily and their specific assignment to stimulus attributes such as "leaf shape" or "bark color" was random for each subject.

In the Control condition, the relative frequency of the different values was the same as in the correlated conditions, but in this group there were no correlations among the values to define distinct categories. Thus, although values 1 and 2 occurred much more frequently than values 3 or 4, as in the correlated conditions, all attributes varied independently over instances so that no attribute value of an instance predicted any of its other values.

As in Experiment 1, the two correlated conditions differed in the order in which the first 24 (out of 36) instances were presented. In the Blocked condition, the first 12 instances were from Category A and the second 12 instances were from Category B. In the Mixed condition, the first 24 instances were presented in an intermixed sequence of As and Bs (i.e., no more than 3 instances from the same category could occur in succession). In both conditions, the first instance shown of both the Category A and Category B blocks had no exceptional values; this ensured that subjects would see the correct default values of all 12 attributes of each category before they saw any exceptional values. The final 12 test trials for all conditions consisted of 6 instances from each category presented in an intermixed sequence.

As in Experiment 1, stimuli were generated from coded specifications from a computer file and the same counterbalancing procedures were used as before.

Results and Discussion

The data collected in Experiment 2 were similar to those of Experiment 1. The accuracy of recognition memory and study-time data from this experiment are displayed in Figure 4.

Insert Figure 4 about here

Pretraining Trials

As in Experiment 1, the data show greater learning of category norms during the pretraining phase (the trials prior to the final 12-instance test block) of the Blocked condition than during corresponding trials in either the Mixed or Control conditions. When subjects' recognition accuracy averaged over all features of the instances is plotted over trials, the Blocked condition shows the same pattern of rapid acquisition for each category as in Experiment 1. Memory for Category A instances increased rapidly (from 0.40 to 0.91) over trials of the Category A block, dropped (to 0.62) when subjects were surprised by the first instance of Category B on the thirteenth trial, but their instance memory increased rapidly again as Category B was learned, reaching about the same asymptotic level (0.94) as for Category A.

Whereas subjects in the Blocked condition appeared to learn rapidly the default values of each category, subjects in the other two conditions showed little evidence of learning during this initial phase. Averaging over the 24 initial trials, overall memory was significantly higher in the Blocked condition than in the Mixed and the Control condition, whereas the latter two conditions did not differ during this phase.

Consistent with this pattern of memory data, during pretraining subjects in the Blocked condition showed longer study-times for exceptional values than did subjects in the other two conditions. Subjects in the Blocked condition studied exceptional values reliably longer (3.02 sec) than default values (2.01 sec). Subjects in the Mixed condition also studied exceptions slightly (and significantly) longer than defaults during pretraining (2.31 vs 2.07 sec). Comparing the two groups, Blocked subjects studied exceptions significantly longer than did the Mixed subjects.

Subjects in the Control condition also showed a marginal tendency ($p < .10$) to study infrequent values (corresponding to exceptions in the other two conditions) longer than frequent values (corresponding to defaults), but they did not exceed the time that Mixed subjects spent studying exceptions.

Although these study-time results were consistent with the recognition memory data described previously, the study-time data, when plotted over trials, exhibited more variability than the smooth recognition learning functions (see Figure 4b). In part, this variability in study-times may reflect the few exceptions per trial plus the fact that exceptions occurred on different attributes on random trials. This positional uncertainty may have caused some subjects to miss (not sample) a hidden exceptional value on any given trial, thus missing the opportunity to study it longer. This factor would increase the variability of the study-time data across subjects and trials within the Blocked condition.

Although subjects in the Blocked condition spent significantly more time studying exceptions than did subjects in the other two groups, memory for exceptional values did not differ significantly among the three groups. (Memory in the Blocked condition exceeded that in the Mixed condition by 0.09 and that in the Control condition by 0.08; these differences failed to attain significance due to the large variability). As expected, within the Blocked condition, memory for exceptions (0.74) was significantly poorer than for defaults (0.87). In contrast, memory for default vs. exceptional values did not differ reliably in either the Mixed or the Control condition. This pattern is consistent with the prediction of greater default learning in Blocked condition compared to the other two.

In addition to being remembered more poorly than defaults, an exceptional value on one trial produced a "carry over" effect in that subjects in the Blocked condition perseverated in studying that (now reinstated) attribute longer on the following trial. On the following instances, subjects in the Blocked condition studied the attribute that had been exceptional earlier for 0.13 seconds longer than average, a marginally significant effect ($p < .10$). This carry-over effect was not significant in either the Mixed or the Control groups (both $p > .10$).

Although encountering an exceptional value caused Blocked subjects to attend more to the same attribute on the following trial, their memory for this default value on that trial actually decreased (by 0.11) compared to their average default memory. This temporarily poorer memory for the default may reflect proactive interference stemming from retrieving the exception-value of the prior instance. This carry-over effect was absent for the other two groups.

The fact that these carry-over effects were larger in the Blocked than the other two conditions seems to indicate that exceptional values were more surprising or unexpected in that condition, as would be predicted if Blocked subjects formed stronger default expectations than did those in the Mixed or Control groups. The existence of such effects suggests that default norms were temporarily reduced in strength when an exceptional value occurred.

Final Test Block

The initial pattern of results, with strong learning in the Blocked condition but hardly any in the other two conditions, was greatly attenuated when instances of the two categories were presented in mixed alternation during the final test block. The memory advantage of the Blocked over the Mixed condition that was present during pretraining disappeared during the test block. Recognition accuracy of Blocked subjects dropped from 0.94 to 0.78 during the test trials, a level that was not significantly higher than that of Mixed subjects. Nonetheless, during the test block both correlated groups continued to reveal evidence of learning when compared to Control subjects whose recognition memory averaged only 0.60 during the test trials. Although recognition of both defaults and exceptions declined somewhat during the test block for the Blocked condition, the difference between them remained significant; on the other hand, this memory difference never approached significance for the Mixed or Control groups.

Consistent with the memory data, study-times for exceptional values indicated a sharp drop (from 3.02 to 2.40 sec) between the pretraining and test phases for Blocked subjects, whereas Mixed subjects remained at the same level (2.47 sec) during the two phases. Both Blocked and Mixed groups showed longer study-times to exceptions during the test block than did the Control group, but the differences fell short of statistical significance. Also, the carry over effects due to an exceptional value that were measurable for both study-times and memory during the pretraining phase of the Blocked condition were reduced to insignificance during the test block.

Implications

The results of this experiment demonstrate that categories can be acquired in an unsupervised task even when the default values of those categories occur with less than perfect reliability. The pattern of results that imply such learning arose most strongly during the pretraining phase for the Blocked condition, but also occurred in attenuated form during the test phase of both the Blocked and Mixed conditions. Learning was indicated by the increased accuracy of recognizing the features of each instance, especially defaults which could be inferred from category norms. Accompanying this improved default memory was a tendency to study exceptional values for a longer time and to show some improvement in memory for these values (although the latter effect was not statistically significant here). In addition, encountering an exceptional value on a given attribute appeared to reduce temporarily the expectedness of its default value, i.e., subjects showed some tendency to increase study-times for default values following an exceptional value of the same attribute in the previous instance, and were less accurate in remembering it due to proactive interference.

Perhaps the most important difference between the results of Experiments 1 and 2 was the attenuation of learning that occurred in the Blocked condition of Experiment 2 when instances were presented in mixed sequence during the final test block. Consistent with the predictions of category invention, initial learning did appear to occur more rapidly in the Blocked than in the Mixed condition. But the fact that learning in the Blocked condition did not exceed that of the Mixed condition during the test block weakens the evidence for probabilistic category invention provided by this experiment.

Recall that the results of Experiment 1 also showed slightly reduced performance during the final mixed test block compared to the earlier single-category pretraining blocks. However, in that case the attenuation did not eliminate the significant differences between conditions as in the present experiment. The attenuation of test-block performance that occurred in both experiments was probably due to the greater uncertainty about the categorization of each instance in a mixed compared to a blocked sequence. This greater uncertainty would make it more difficult for subjects to remember the category membership of each instance, forcing them to spend some time rehearsing this categorization during the study period, presumably at the expense of focusing specifically on the non-default features of each instance. This added memory demand could lead to reduced learning effects in both the memory and the study-time measure.

The fact that the attenuation in the present experiment eliminated the test-trials difference between the Blocked and Mixed condition, whereas these differences remained significant in Experiment 1, is presumably related to differences in the types of informative values used in these two experiments. The frequent occurrence of exceptional values in Experiment 2 may have weakened subjects' category norms, making them more vulnerable to mutual interference and confusion. Thus, subjects in the Blocked condition of Experiment 2 might have suffered significant forgetting of their Category A defaults during the block of Category B instances. In turn, this forgetting of the Category A pattern may have resulted in decreased performance and confusion between categories in the mixed test block.

The present results raise the possibility that, even after subjects learn to distinguish separate categories, they may still have difficulty keeping straight all their elements. Thus, while category invention would enable subjects to capture correlational structure relatively efficiently without having to learn a large matrix (22 x 22 here) of interfeature co-occurrences, such categories still appear to be somewhat subject to interference from related categories. Analogously, although subjects in standard verbal learning experiments are quite aware that they are learning two or more distinct lists, akin to different categories in the present experiment, they still show appreciable interference and forgetting due to learning multiple lists (see, e.g., Postman, 1971; Millward, 1971).

Experiment 3

This experiment aimed to provide further evidence for category invention. The invention process depends strongly upon subjects perceiving a surprising contrast -- a novel instance that breaks strong expectations. By comparison, an autocorrelation approach emphasizes simple frequency of experiencing co-occurring features as important for accumulating data regarding their correlation. In Experiment 3, we pit the "contrast" manipulation against frequency-of-covariation to see which is more potent in promoting the acquisition of categories in an unsupervised environment.

In the *Contrast* condition of Experiment 3, subjects were presented with a long series (16) of Category A instances before seeing a mixed test block of A and B instances. These subjects should learn strong defaults for Category A during this pretraining, so that they should be likely to invent a second category when surprised by the first B-instance. Moreover, they should be able to maintain the separate integrity of the first-learned Category A norms as they respond to the mixed series of A's and B's.

The Contrast condition was compared to a *Practice* condition, in which the pretraining consisted of an equal number of instances (8) from both categories, presented in mixed sequence, after which they saw the same mixed test block as did the Contrast subjects. According to a frequency-of-covariation account, the Practice subjects should exceed the Contrast subjects in learning the feature correlations underlying the B category, since they will have seen 8 exemplars of these correlations during the first phase of the experiment, compared to none for the Contrast condition. Similarly, having fewer A instances in the first phase should produce less interference (than in the Contrast condition) for the Practice subjects to learn the correlations underlying the B category.

In contrast to these predictions, the category invention hypothesis predicts that the early mixing of A's and B's should mislead subjects into assimilating them to a single, overly general category so that they fail to notice and record contingent covariations. The early mixing of As and Bs makes it difficult to discriminate default from variable attributes. On the other hand, the novel surprise arranged for the Contrast subjects should trigger their invention of a B category distinct from their A category, and thus provide a basis for their noticing and recording contingent covariation of features using the two categories. It is this separate maintenance of category norms that then permits the Contrast subjects to respond appropriately to instances of Category A vs. Category B during the final block of trials.

In sum, category invention expects better learning of Category B in the Contrast condition, even though the only difference between the Practice and Control conditions is that in the former, eight instances of Category B were replaced with instances of Category A. Such a contrast effect would indicate the importance of subjects' perception of contrast or mismatch between categories in their unsupervised learning. Such results would falsify the prediction of an autocorrelation approach that experts, all else being equal, learning of correlations should increase with practice.

Method

Subjects, Materials, and Procedure

The subjects were 31 students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement. The experimental procedure was identical to that of Experiment 1 and consisted of 40 trials plus instructions and debriefing. The descriptions of tree instances were designed according to the same specifications used in Experiment 1 and the training instances were partitioned into the same two categories based on correlations among nine of the twelve stimulus attributes.

Design

Subjects were randomly assigned to two conditions differing only in the sequencing of the training instances. In the Contrast condition, instances of Category A were presented for the first sixteen trials, referred to as the pretraining block. Following this pretraining, a mixed test block of twelve instances of each category occurred in random order (randomized for each subject). In the Practice condition, the pretraining block consisted of eight A-instances and eight B-instances presented mixed together in a random order. After pretraining, they then had the same test block of 12 As and 12 Bs as did the Contrast subjects.

In both conditions, instances were so constructed that all four values of each variable attribute occurred an equal number of times within each category; within this constraint, values of these attributes were assigned randomly. The same stimulus set was presented to all subjects in a given condition, but the order of specific instances within the pretraining and test blocks was randomized anew for each subject.

Results and Discussion

The same types of data were collected in this experiment as in Experiment 1. The average study-times and recognition-memory accuracies are displayed in Figure 5. The main prediction of the category invention theory is that strong learning would occur in the Contrast condition while relatively little learning would occur in the Practice condition.

Insert Figure 5 about here

Contrast Condition: Recognition Memory

Consistent with these expectations, both recognition memory and study-time data showed evidence of significant learning in the Contrast condition. Turning first to the recognition memory data averaged over all 40 trials of the experiment, Contrast subjects remembered defaults with a mean accuracy of 0.94, which was significantly higher than the 0.84 for variables. Learning occurred rapidly during the pretraining A-block, with recognition memory increasing from 0.48 on the first trial to 0.88 on the eighth trial (see Figure 5a). Thereafter, memory for Category A instances remained high and stable for the remainder of the pretraining and throughout the test block.

Accuracy decreased sharply on the first B-trial of the test block, compared to the preceding A-trial. Thereafter, accuracy increased from 0.68 to about 0.93 and remained stable thereafter. Asymptotic accuracy of Category B was about equal to that of Category A. Thus, prior learning of Category A appeared not to impair learning of Category B, as would have been expected by autocorrelation models that include associative interference.

Contrast Condition: Study-Time

The study-time data showed a pattern of rapid learning in the Contrast condition similar to that shown by the memory data (Figure 5b). Over all initial trials, variable attributes were studied 1.33 sec longer than defaults. During pretraining, the preference index increased significantly from -0.16 on the first trial to 2.08 sec on the sixteenth trial.

Preference scores dropped significantly on the first B-instancet trial compared to the preceding A-trial (2.08 sec vs -0.19 sec). The -0.19 score means that subjects regarded the new defaults of the B-category as highly informative on that trial, and therefore studied those novel defaults slightly longer than the variable attributes. However, preference for examining variable attributes increased rapidly over the twelve B-instances in the test block, implying rapid learning of the B-norms. Final learning of Category B equaled that of Category A.

Overall, the Contrast condition showed strong learning of Category A during the pretraining block, no significant reduction of this A-learning during the mixed test block, and strong B-learning during the test block. Since final learning of Category B did not differ from that of Category A, it appears that there was no significant interference between the categories in this condition.

The Practice Condition

The Practice condition was similar to the Mixed condition of Experiment 1, and, as before, produced little evidence of significant learning. Recognition accuracy was significantly greater for defaults (0.72) than for variables (0.64). Variables were studied slightly longer than defaults in this condition, but insignificantly so. Since the study-time data showed little evidence of learning by subjects in this condition, their greater accuracy in verifying defaults compared to variables was probably due mainly to better guessing of the correct values of defaults (which presented only two values within the training instances) than of variable attributes (which had four values presented).

Comparing the Two Conditions

Direct comparisons of recognition memory between the two groups (see Figure 5a) supported the conclusion that significant category learning occurred in the Contrast condition but not in the Practice condition. Instance memory for the Practice subjects was essentially constant from Trial 2 onwards. Overall memory for instances was significantly greater in the Contrast condition during the test block, for both Category A and B.

The difference in learning between the Contrast and Practice conditions was also supported by comparisons of the study-time data from the two groups (see Figure 5b). The mean study-time preference for studying variables was 1.33 sec in the Contrast condition, which was significantly greater than the corresponding 0.23 sec preference in the Practice condition. This comparison was significant for both categories and remained so when examining only the final test block (which was identical in both conditions).

In addition, recognition-memory was greater in the Contrast condition for both defaults and variables. Contrast subjects' better learning of variable attributes (compared to Practice subjects) is consistent with their greater preference for studying variables. The result suggests that Contrast subjects used their category knowledge to improve their learning of both predictable and unpredictable features of the instances.

The finding that Category B was learned better in the Contrast than the Practice condition was the basic "contrast effect" we were seeking. An autocorrelation process would seem strained to accommodate this apparently paradoxical finding, that decreasing the number of instances studied from a given category (i.e., Category B) increases later learning of that category. A strictly autocorrelational approach also would have expected some interference in learning the feature correlations in the Contrast condition; yet none seems to have occurred. Thus, this contrast effect provides strong evidence that subjects used category invention to capture the correlational patterns in the present experiment.

Experiment 4

The purpose of this experiment was to replicate the contrast effect obtained in Experiment 3 and extend it to categories based on *probabilistic*, rather than deterministic, correlational patterns. While Experiments 1 and 3 both provided evidence for category invention in the deterministic case, Experiment 2 fell short of providing definitive evidence either for or against category invention in the case of probabilistic categories. Therefore, the present experiment, if successful, would be the first to provide strong evidence for the use of category invention to learn probabilistic patterns in an unsupervised environment.

Method

Subjects, Materials, and Procedure

The subjects were 35 students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement. The experimental procedure was identical to that of Experiment 2, and consisted of 40 trials plus instructions and debriefing. The same stimuli (list descriptions of fictitious trees) were used as before, and these were divided into the same correlation-based categories. As in Experiment 2, one-quarter of the instances within each category had no exceptional values, one-half had a single exceptional value, and the remaining one-quarter had 2 exceptional values. Exceptional values occurred equally often on all 12 attributes, and different exceptional values were used in the two categories (e.g., value #3 for Category A and value #4 for Category B).

Design

Subjects were randomly assigned to two different conditions, analogous to those of Experiment 3. In the *Contrast* condition, the first 16 of 40 instances were all members of Category A. Following this pretraining block, the next 24 instances consisted of 12 As and 12 Bs, presented in a mixed sequence. No more than 3 successive instances from the same category occurred during this test block. The first instance shown of both Category A and Category B had no exceptional values, allowing subjects to see the correct default values of each attribute prior to encountering any exceptional values.

Subjects in the *Practice* condition saw the same 24 instance test block as those in the Contrast condition, but their pretraining block consisted of 8 instances of Category A and 8 instances of Category B presented in Mixed sequence. No more than 3 successive instances of the same category occurred during either the pretraining or test phase. The same procedures for balancing the experimental design were undertaken as in previous experiments.

Results and Discussion

The same types of data were collected in this experiment as in Experiment 2 and are shown in Figure 6. The major result predicted by category invention was that learning should be greater in the Contrast condition than in the Practice condition; importantly, this should occur for both the pretrained category (Category A) and the second category (Category B).

Insert Figure 6 about here

Recognition Memory

Analyses of recognition memory pooled across defaults and exceptions provided strong support for the contrast prediction (Figure 6a). The pattern of recognition results was similar to that of Experiment 3. During pretraining, because Contrast subjects received 16 instances of only one category, whereas Practice condition subjects received 8 A's and 8 B's, both theories expect higher instance memory at this point in the Contrast condition than in the Practice conditions. As expected, Contrast subjects indeed showed higher instance memory during these trials.

However, when instances of Category A and Category B were mixed together in the final test block, Contrast subjects continued to show better instance memory than did Practice subjects (0.83 vs. 0.67, respectively). Importantly, this group difference held true not only for Category A, but also for Category B ($p < .05$).

During the test mixed block of the Contrast condition, memory for Category A was reliably higher (by 0.08) than memory for Category B. This difference was probably due to subjects receiving the 16 pretraining exposures to Category A before seeing any Category B instances.

Although subjects in both conditions studied exceptional values for longer than default values (see below), memory for default values was higher than that of exceptions in both conditions. Contrast subjects' memory for defaults in the test block was 0.85, compared to 0.69 for exceptions; for Practice subjects, memory for defaults was 0.68 compared to 0.61 for exceptions ($p < .02$). While this latter result suggests some default learning in the Practice condition, the 0.16 difference in the Contrast condition was significantly greater than the corresponding 0.07 difference in the Practice condition ($p <$

.05) -- a result mainly due to Contrast subjects' high memory for defaults.

This difference in Category B learning was the primary contrast effect for which we were searching. An autocorrelation process cannot accommodate such a contrast effect, since reducing the instances from Category B and substituting potentially interfering instances of Category A significantly improved later learning of Category B. However, the outcome is explained by the category invention hypothesis: learning strong default norms about one category results in heightened contrast upon being exposed to the first instances of a second category.

Study-time Data

The study-time results (see Figure 6b) were generally weaker than the recognition data, and revealed fewer interesting differences between conditions. Exceptional values were studied longer than default values in both conditions; this effect appeared during pretraining and remained significant throughout the experiment. However, only during pretraining did study-times to exceptional values in the Contrast condition significantly exceed those in the Practice condition.

Recall that in Experiment 2, the occurrence of an exceptional value caused subjects to increase their study-times to that attribute of the following instances. No such carry-over effects in study-time attained significance in the data from the present experiment. However, memory for the reinstated default occurring on the trial following an exception was significantly reduced ($p < .05$) in the pretraining block of the Contrast condition, as happened in Experiment 2.

Implications

Taken together, the study-time and recognition memory results imply that both categories were learned better in the Contrast than in the Practice condition. Thus, the data provide another example of improving learning of Category B by replacing its instances during pretraining with instances of Category A. This result violates a prediction of a pure autocorrelation process, namely, that learning of a given correlational pattern should increase with the number of instances encountered from that pattern, and that exposure to an instance of a given category should improve learning of that category by at least as much as exposure to an instance of a different category. Rather, it appears that our subjects were learning the patterns by inventing a discrete category in response to the mismatch or perceived contrast between instances of Category B and previously acquired norms about Category A.

Recall that in Experiment 2, involving a similar comparison between a blocked and mixed training sequence, no significant difference between conditions was observed when instances of the two probabilistic categories were intermixed during the test block. This raises the question of why the partially blocked sequence used in the present experiment produced better learning, relative to a mixed sequence condition, than did the fully blocked sequence used in Experment 2. A plausible explanation is suggested by considering that subjects in Experiment 2 saw 12 instances of Category B in succession before returning to another instance of Category A, whereas in Experiment 4 subjects continued to see instances of Category A interspersed among Category B trials. Subjects in Experiment 2 may have forgotten some of their Category A learning while acquiring Category B, due either to decay or interference. This forgetting of Category A norms would have caused subjects to become confused when instances of Category A were later mixed with instances of Category B in the test block, affecting their encoding and memory performance for instances of both categories.

The results of this experiment, together with those of Experiment 2, demonstrate that people can learn without supervision categories based on probabilistic correlations, and that the instance-memory task can provide an index of such learning. In addition, the results provide evidence for category invention in learning probabilistic co-occurrence patterns similar to that provided by Experiments 1 and 3 for deterministic patterns. However, the memory measure proved a more reliable index than did the study-time measure for the two experiments concerned with probabilistic patterns, whereas both measures provided equally valid indices when subjects learned categories based on deterministic patterns.

Why would subjects' tendency to focus on non-default values be stronger when those non-defaults are routine variables, as in Experiments 1 and 3, than when they are exceptions that directly violate default expectations, as in Experiments 2 and 4? If the attentional salience of a given attribute value varied only with its improbability within a category, default violations should have greater salience and show greater study-time effects than routine variables. However, we would propose that the attentional salience of a given attribute value depends not only on the improbability or unexpectedness of that specific value, but also on the average salience or informativeness of the attribute to which it belongs. Thus, routinely variable attributes were informative (had non-default values) for every instance of Experiments 1 and 3, and thus subjects would have learned to attend to these attributes consistently. By contrast, exceptional values in Experiments 2 and 4 occurred within attributes that nearly always had default values, so these attributes usually conveyed relatively little information. Perhaps the greater improbability of the exception values was offset by the lower *a priori* salience of the attributes within which they occurred. In other words, the fact that subjects in Experiments 1 and 3 could learn to attend consistently to specific variable attributes on every trial resulted in a more durable preference for those attributes. On the other hand, subjects in Experiments 2 and 4, for whom exceptional values could occur on any attribute on any trial, could not bias their attention in advance towards attributes with exceptional values.

Experiment 5

Experiment 5 was similar to Experiment 4 in most respects, comparing Contrast and Practice conditions, but in order to strengthen our conclusions the training stimuli contained fewer default violations than did those of Experiment 4. Rare exceptions should thus appear more informative when they do occur, and should be less disruptive of subjects' category norms and their procedures for encoding individual instances. The greater surprisingness of individual exceptions combined with the greater predictability of the instances should combine to increase subjects' tendency to focus on exceptional values. As a result, subjects should be more likely to retain strong category norms during a mixed test sequence and reveal greater evidence of learning on both the recognition memory and study-time measures.

An uncorrelated Control group was included in the present experiment to provide a baseline to evaluate any learning observed in the other two groups. Thus, even if the Contrast and Practice conditions did not differ, it might still be possible to conclude that learning had occurred in these groups by comparing them to the Control group.

Method

Subjects and Procedures

The subjects were 36 students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement. The same procedure was employed as in previous experiments, and the same stimulus materials were employed. The experiment consisted of 48 study-test trials, and subjects were allowed 90 min to finish.

Subjects were randomly assigned to three groups; two of these, the *Contrast* and *Practice* conditions, were similar to those of Experiment 4. Within these conditions, default attribute values were correlated probabilistically, as in Experiments 2 and 4, but the correlational patterns of the present experiment were more reliable. Over the experiment as a whole, one half of the training instances from each category had a single exceptional value and the other half contained all default values. The overall probability of default values in this experiment was thus approximately 0.96.

As before, the Contrast and Practice conditions differed in whether instances of one or both categories were presented during 16 pretraining trials. The same mixed test block of 16 A's and B's was employed in both groups. The first half of the pretraining showed instances with no exceptional values, as did the first four trials of the test block.

A Control condition similar to those of Experiments 1 and 2 was pretrained before receiving the same test block as the other two conditions. However, their 16 pretraining instances lacked correlated values and distinct categories. During their pretraining, the frequency of each specific attribute value was the same as during the corresponding trials of the Practice condition, and exceptional values occurred on the same trials in both conditions.

Instances were presented in the same order to all subjects within a given condition, but instances were constructed from abstract numerical codes whose correspondence to actual stimulus attributes was randomly decided for each subject. The same procedures for balancing the experimental design were used as before, i.e., presenting attributes in a different order to each subject, random order of testing attributes, and so on.

Results and Discussion

Experiment 5 had two primary goals: (1) to provide further evidence for the contrast effect and discrete category invention, and (2) to demonstrate that learning of categories based on probabilistic correlations can be indexed by study-times as well as by recognition memory accuracies. The same types of data were collected as in previous experiments, and are shown in Figure 7.

Insert Figure 7 about here

Recognition Memory Data

Considering first the recognition data pooled across default and exceptional values, the pattern of results in the present experiment replicates that of Experiment 4 (Figure 7a). Memory in the Contrast condition was higher than in either of the other two conditions throughout the experiment. During the final test block, memory averaged 94% in the Contrast condition, 76% in the Practice condition, and 73% in the Control condition. The Contrast condition reliably exceeded the other two, which did not differ. Thus, the memory data suggests that significant category learning occurred only in the Contrast condition, with little evidence of category learning in the Practice condition.

To provide strong evidence for category invention, learning should be greater in the Contrast condition than the Practice condition for both the pretrained category (A) and the non-pretrained category (B). This group difference was indeed significant at the .02 level for Category A and at the .05 level for Category B. The latter result was the basic "contrast effect" we were seeking in this experiment.

The results of Experiment 5 differed somewhat from those of Experiment 4 in that only the Contrast subjects showed higher memory for defaults compared to exceptions. Within the test block, this difference was significant at the .02 level. On the other hand, Practice and Control subjects remembered defaults and exceptions at the same rate.

During the test block, accuracy of remembering exceptions averaged 80% in the Contrast condition, 78 % in the Practice condition, and 68% in the Control condition. The latter is not significantly below the first two groups.

Study-time Data

The study-time data provided evidence of category learning in both the Contrast and Practice conditions (Figure 7b). During the test block, Contrast subjects studied exceptions significantly longer than defaults. This difference was also significant in both the Practice condition and the Control condition over the same trials. Exceptions were studied an average of 3.88 sec in the test block of the Contrast condition, 3.90 sec in the Practice condition, and 2.66 sec in the Control condition (compared to average default study-times of approximately 2 sec). Due to differences in variability, the Contrast subjects but not the Practice subjects reliably exceeded the Controls in time spent studying exceptions. The pattern of study-time results seem to imply that some learning may have occurred in the Practice condition, which is puzzling since their memory data showed very little evidence of learning. It is possible that the longer study-times for exceptional values observed in the Practice condition is partly spurious, since one outlier subject contributed greatly to the high study-times for exceptions observed in that condition.

These results indicate that unsupervised learning of categories with probabilistic defaults can be indexed by both study-times and recognition memory, similar to the learning of deterministic categories. However, the study-time measure is apparently more vulnerable to the confusion that may occur when defaults are violated too frequently, as presumably occurred in Experiments 2 and 4. Reducing the number of exceptions in the present experiment apparently increased the saliance of these values when they did occur, in addition to reducing the disrupting effects of such violations on subjects' default norms for the categories. As a result, significant category learning effects were revealed by the study-time index in this experiment.

General Discussion

These experiments had three primary objectives: (1) developing and testing the instance memory task as a procedure for investigating unsupervised learning; (2) providing evidence regarding category invention in unsupervised learning; and (3) establishing the generality of the results for new verbal categories based on both deterministic and probabilistic covariance patterns.

These objectives were largely achieved by the experiments described above. Experiments 1 and 3 demonstrated unsupervised learning of category norms based on perfectly correlated features, and this learning was clearly indexed by both study-times and recognition memory data. These experiments also suggested that subjects learned the correlational patterns by hypothesizing discrete categories. The experiments confirmed and extended earlier results implicating category invention based on deterministic correlational patterns (Clapper & Bower, 1994). Evidence for category invention has thus been obtained with two different stimulus types (verbal stimuli in the current experiments, pictorial stimuli in Clapper & Bower, 1994), and with two very different tasks (free listing of stimulus attributes vs. encoding times and recognition accuracies from the instance memory task). Taken together, these experiments provide converging evidence for the generality of category invention as an important component of unsupervised learning in humans.

Experiments 2, 4, and 5 provided demonstrations of unsupervised learning of categories based on less than perfect (probabilistic) feature covariances. The results suggested that subjects captured such probabilistic structures by inventing categories much as they did in the deterministic case. The evidence for category invention derived mainly from significant "contrast effects" in recognition memory, as reported in Experiments 4 and 5.

These results do not eliminate the possibility that subjects in these experiments also learned some inter-feature correlations. For example, the small learning effects observed in the Mixed or Practice conditions of our experiments may reflect the gradual strengthening of interfeature associations (see Clapper & Bower, 1994). However, it is clear that autoassociation alone cannot account for the strong contrast effects observed in these experiments. Thus, although the present results provide strong demonstrations of category invention, we do not claim that autoassociative learning never occurs in this setting.

At what level of unreliability would a correlational pattern become essentially unlearnable by human subjects in an unsupervised environment? Consider, for example, the probabilistic categories employed in many supervised category learning experiments, in which diagnostic features might predict a particular category with, say, probability .75 and in which, even at asymptotic levels of learning, the maximum possible accuracy of classification is barely above chance (see, e.g., Estes, Campbell, Hatsopoulus, & Hurwitz, 1989; Homa, 1984; Medin & Schaffer, 1978). It seem likely that such categories would be practically unlearnable by subjects in unsupervised tasks. So how could children learn real-world fuzzy categories without explicit instruction?

One possibility is that the actual fuzzy categories that children are able to learn easily without supervision tend to be distinguished from related categories by highly salient features that are present with high reliability and are only rarely violated (e.g., animals that fly are usually birds). At least, salient features seem to characterize many biological categories, and serve as a basis for many published identification guides (e.g., bird-watchers' books) that catalogue such categories in terms of their distinguishing features. For example, features that distinguish different species of trees or birds are often quite reliable. On the other hand, when such reliable distinguishing features are absent, people may be unlikely to discover such categories spontaneously without some form of supervised tutoring.

Alternative Clustering Proposals

As noted in the Introduction, a number of techniques for conducting conceptual clustering have been proposed in the areas of numerical taxonomy, psychological scaling, and artificial intelligence (Everitt, 1980); therefore, we will briefly review several proposals to see whether they might apply to our experiments and explain our results.

Most of the clustering proposals have not been formulated as incremental learning procedures; in fact, it is commonly assumed that all *n* patterns to be partitioned are available for simultaneous inspection and grouping. Moreover, the typical goal of the proposals is simply to group patterns into clusters that are "optimal" according to one or another criterion. As a result, we have encountered considerable difficulty applying these ideas to our experimental methods, especially to our particular dependent measures of clustering. It is therefore difficult to decide what predictions the various techniques make regarding the "blocking vs. mixing" variable manipulated in our experiments.

A number of psychological scaling techniques for clustering *n* objects (patterns, instances) begin with the complete *n x n* matrix of inter-object similarities. Clustering methods are then applied either directly to the similarities or to the inter-object distances derived from multidimensional scaling of the similarity matrix (Carroll, 1976; Shepard, 1962; Torgerson, 1952). The scaling solution can be either "metric" if the similarities are interpreted as ratio or interval measurements, or "nonmetric" if the solution is derived from only the rank ordering of the similarities.

In some methods, clustering proceeds by designating in advance the desired number of clusters (categories). The clusters are then composed by putting together those objects that are highly similar to one another. One criterion driving the process is to search for "good clusters" that maximize the average similarity among instances within a cluster. A secondary criterion would be to also maximize the average distance (dissimilarity) between clusters.

Other methods do not begin with presuppositions about the number of categories; rather, they proceed to form whatever clusters the data suggest within the constraints of optimizing average within-category similarity. Of course, using this criterion alone, the optimal partitioning is the degenerate one formed by assigning each object to its own unique cluster. Consequently, further criteria regarding minimizing the number of clusters and between-cluster similarity are typically added to lead to more interesting partitions.

A variety of hierarchical clustering techniques have also been proposed which aggregate objects that are most similar (e.g., Johnson, 1967; Sattah & Tversky, 1977; Shepard, 1980). The schemes follow an agglomerative, "bottom up" procedure of aggregating close objects together into groups, then aggregating additional objects into existing groups, then aggregating groups together repeatedly into successively larger groups. The methods thus construct a hierarchy (or "dendogram") in which interobject distance is reflected in the number of tree-links connecting the two objects.

Regrettably, these clustering methods appear rather inapplicable as descriptions of incremental learners. In our incremental task, subjects must rely on their impoverished memory, and they have nothing remotely resembling the full *n x n* similarity matrix to use to calculate optimal clusters. Without perfect instance memory (for our 12-attribute stimuli), it is doubtful whether our subjects would be able to derive or retain the matrix of similarities that these methods require. But even if some means could be provided to enable subjects to have available the matrix of interobject distances, the approach implies that the clustering solution is (or should be) independent of the order in which the objects are inspected during training. It is just this implication that our results seriously dispute.

Moreover, there is no obvious way to relate a subject's internal dendogram to our behavioral measures, such as feature study-time, feature recall, and attribute listing. Certainly, if branches of dendograms were divided by alternative values of default features, with instances differentiated at successively lower levels of the net, and if the dendogram in memory were to be entered from the top node and searched downwards (as most net models do), the model would predict most inspection of default features and least inspection of variable features. But this is just the reverse of the inspection behavior observed.

It may be unfair to view such scaling procedures as presumed descriptions of actual processes by which subjects encode and remember the patterns. In fact, scaling theorists usually confine their statements to the claim that the scaling solution is just one *representation of similarity data*, not itself a behavioral process theory.

Clustering Algorithms from Artificial Intelligence

Several clustering algorithms have also been proposed in the literature of artificial intelligence. Their goal is typically to find the best way to group together collections of examples to optimize some function. For example, Hanson and Bauer's (1989) WITT program computes correlations between all possible feature pairs in the collection of instances. Using these, WITT then follows an algorithm that seeks to maximize the average pairwise featural correlations of objects within categories while minimizing the average featural correlations between all contrasting categories. For the stimuli used in our usual experiment (12 four-valued attributes), WITT would have to keep track of 66 contingency tables with 4 to 16 entries per table -- all of which would seem a bit much for our subjects who show immediate recognition-memory for only about half of the attributes of a just-studied pattern (prior to learning category defaults). Furthermore, contingency tables are insensitive to the order in which the entries come in, so the model would expect little influence of the blocking vs. mixed order of instances studied in our experiments.

Another class of AI clustering models build hierarchical trees incrementally, similar to the discrimination nets of the Elementary Perceiver and Memorizer (EPAM) theory (Feigenbaum, 1963; Richman, 1991; Richman & Simon, 1989). Examples include Kolodner's (1983) CYRUS, Lebowitz's (1982) UNIMEM, and Fisher and Langley's (1990) COBWEB. These models typically assume that the person stores and retains a complete record of each instance. Instances are used to grow a discrimination net by adding branches in the tree corresponding to "important" differences between the current instance and the stored replica of an earlier instance sorted down to that same node along branches of the tree. The models differ slightly in what events trigger growth of new branches in the discrimination net. For example, COBWEB divides a node into two branches (two subcategories) by using a "category utility" heuristic (Gluck & Corter, 1985). Category utility refers to how much one can improve (above baseline) the predictability of features of an instance by being told its category. COBWEB subdivides a given branch (cluster) if by so doing it significantly increases the average category utility over the entire set of current categories.

While these net-growing systems are indeed sensitive to the order in which the examples are shown, they apparently predict (P. Langley, personal communication, August, 1994) that best learning in our experiments would occur for the mixed rather than the blocked series of category exemplars. The nets grown by the models would clearly reflect more accurately the logical structure of the instances and be organized more simply after encountering a mixed sequence of instances rather than a blocked sequence, and this net complexity would be reflected in the model's performance. Since these models store all features of all examples (e.g., enabling calculation of category utility to guide the search), it also is not clear how the models would be modified to deal with impoverished memory or to apply to our tasks of attribute listing, measured inspection times, and instance recognition memory.

In summary, our brief survey of this clustering literature has turned up relatively little of direct applicability to our task. In fairness to those models, however, we must emphasize that their goals were typically very different from ours. For example, many AI models are primarily concerned with how to characterize optimal clusters in non-obvious domains, and how to compose algorithms that will converge upon nearly optimal clusters given large collections of noisy instances. In contrast, the categories we studied are exceedingly simple and regular -- defined by simple conjunction -- and all the AI theories would discover them almost immediately, indeed in trivial time. Our goal rather has been to study the way subjects process early instances without external memory aids to help them keep track of prior instances. Clearly, our memory-limited subjects find the incremental learning task difficult despite the logical simplicity of the underlying categories.

Exemplar Storage Models

A class of popular models for category learning tasks are those that assume that complete exemplars are stored, and that a new instance is classified depending on its similarity to the various categories of stored exemplars (Estes, 1994; Medin & Schaffer, 1978; Nosofsky, 1988). It is not clear that the exemplar storage model has the mechanisms required to deal with the performance measures we collect in our unsupervised learning tasks. Our subjects are asked to memorize instances, not classify them. The basic exemplar model has no mechanism by which it can use past knowledge to direct attention preferentially to variable rather than the default attributes of instances. Moreover, the rules that have been proposed (e.g., Nosofsky, 1984; 1986) assign greater salience and attention to the category-defining (default) attributes rather than the variable attributes. To deal with our selective encoding data, the exemplar model would require the addition of some rules essentially equivalent to the "schema plus corrections" strategy. For example, given a value of 1 on attribute 1 in Table 1a, the rule would automatically fill in 1's for default attributes 2-5 and then devote study-time to encoding the presented values of variable attributes 6-8. But such a rule essentially concedes the argument to schema theories. Using one default to fill in others is the defining feature of "property inheritance" that is central to schema theories. Therefore, such a rule would violate the cardinal assumption of exemplar models, viz., that entire exemplars are to be stored, not filled in with default values.

Andersons's Rational Model

An alternative that holds promise of accounting for our data is the rational theory of categorization proposed by J. R. Anderson (1991). This model is a contender because it uses a discrete category invention process to deal with successive instances that are greatly mismatched. The model operates equally well with supervised and unsupervised learning environments.

The rational theory sets up a category for the first instance it sees, and then decides for any later instance whether to include it in an existing category or to set up a new category for it. Using sa priori assumptions, the model attempts to estimate the Bayesian probabilities of the current instance given either of these two states of affairs (previous or new category), and it selects that decision that maximizes the probability of the present instance. Calculation of the Baysian probabilities requires prior assumptions regarding the number of yet-to-be-seen values of features, the weighting of new data against prior assumptions regarding feature probabilities, and a "coupling parameter", c, which is the prior probability that any two objects of the domain will belong to the same category.

In general, the rational model is sensitive to the sequence in which instances are encountered. However, given the very marked differences between our A and B categories (see Figure 1a), the rational model predicts that the two categories will be learned very rapidly with either the mixed or the blocked presentation sequence. For most values of the c parameter (Anderson uses .30 for most of his simulations), the rational model would predict either no difference in category learning due to blocking, or a slight advantage due to mixing instances of the two categories early in training (J. R. Anderson, personal communication, August 1994). The model may possibly produce a slight advantage for the blocked over the mixed presentation condition, but only by assuming implausible values of the parameters (J. R. Anderson, personal communication, August 1994).

The rational model, like many of the other models discussed previously, assumes perfect memory for the instances encountered on each trial. However, this unrealistic assumption may be weakened simply by assuming that some of the feature-to-category probability adjustments that occur on each trial are lost or forgotten by the following trial. Such forgetting would tend to reduce the perceived difference between instances of the A and B categories early in a mixed sequence, and increase the chance of lumping them together into a single category. We suspect that, with this modification, the rational model could simulate our results under a somewhat wider range of parameter values.

To relate the rational model to our performance measures, we might assume that once an instance is categorized using one default attribute, the model-subject uses the conditional norms of that category to "fill in" the rest of the default values, and then spends most of the remaining time encoding the variable features of the instance. In that manner, the model would mimic a "schema plus corrections" encoding strategy. Because the rational model has the potential for explaining our results, we consider it to be a plausible description of the discrete category-invention process. An advantage of linking our results with the rational model is that it has a proven track-record for explaining the standard results in the categorization literature (see J. R. Anderson, 1991).

Implications for Memory Performance

Beyond providing evidence of how categories are induced in unsupervised learning tasks, the present experiments also relate to the issue of how schematic knowledge about categories affects people's episodic memory for instances. As noted in the Introduction, research areas that bear on this topic include people's memory for text passages based on stereotyped routines or scripts (e.g., Graesser, Woll, Kowalski, & Smith, 1980; Bower, Black & Turner, 1979; Schank & Abelson, 1977), memory for descriptions of people based on personality stereotypes (e.g., Srull & Wyer, 1989), and differences in memory between domain experts and novices (Chase & Simon, 1973; DeGroot, 1965, 1966). Most of this research assumes an encoding strategy based on a "schema plus corrections" strategy (see, e.g., Clapper & Bower, 1991; Graesser et al., 1980; Schank and Abelson, 1977); that is, instances are encoded by referring or "pointing" to the appropriate schema (set of category norms) in long-term memory, and then selectively encoding any features of the instance that could not be predicted from this schema. We assumed that subjects would employ such an encoding strategy in the present experiments, and we used their tendency to do so as an index of how well categories were learned.

The schema-plus-corrections encoding strategy makes several predictions about memory performance; we will mention three. First, memory should be better when a schema is available than if one is not (a common difference between experts and novices, for example). Because the schema specifies many feature-values of the instance in advance, it reduces the amount of new information that must be encoded to remember the instance accurately. Second, the theory predicts that, in the case of stereotypic text passages, subjects will show better discrimination in recognition memory for atypical or unexpected values of default features because they pay more attention to non-defaults while encoding the passages (see Heit, 1993).

A third implication of the schema-plus-corrections encoding process is that as successive instances of a category are presented, their consistent features will become more predictable and so receive progressively less attention, freeing the subject to devote more time to learning the unpredictable features of later instances. The present experiments allowed us to observe the time-course of this learning over trials. Our method also allows us to provide a more detailed description of the factors that determine attentional allocation, thus refining the analysis of simple dichotomies such as that between "typical" and "atypical" features. For example, we suggest that attention may be controlled not only by the improbability or unexpectedness of a particular value of an attribute, but also by the average informativeness or utility of that attribute as determined over previous trials. Thus, even though routine values of variable attributes may occur more frequently and so appear less informative or surprising than exceptional values, the fact that a variable attribute has proven consistently informative over previous trials will cause it to be attended to more often in the future. In other words, rather than attention and encoding being controlled by a single factor within a category (expectedness or typicality); a two-factor model stressing the informativeness of a given value and the informativeness of the overall attribute is more appropriate.

A final point relevant to the schema-plus-correction theory concerns the low recognition memory discrimination usually found with default features (e.g., Graesser et al., 1980; Bower et al., 1979; Heit, 1993). We note that such results are often obtained in experiments which describe events in narratives. By convention, narrative descriptions are highly abbreviated and failure to mention an expected feature does not imply its absence from the situation. However, in cases in which the *absence* of a default would be noticed by subjects as an explicit violation of their expectations, we would expect high overall memory discrimination for defaults even if subjects generally gave little attention to them when they were present. This outcome held true for the present experiments in which exceptional values were substituted for defaults and were explicitly noticed by subjects as exceptional.

In earlier work, we showed that subjects will notice the *absence* of additive features of a stimulus when these are strongly expected; they will then count this absence as a salient characteristic of the stimulus. For example, Clapper and Bower (1991) showed that when subjects strongly expected instances within a category to have some feature (such as wings on a fly), the absence of that feature was readily noticed (e.g., a fly without wings) and this missing value was more salient than when the default value was present. In such cases, we would not expect subjects to suffer the same low memory discrimination for defaults that is often observed in experiments using text materials.

Final Comment

The study-time task investigated here provides results that converge upon the same conclusions about category invention that we arrived at with our earlier investigations using the attribute listing task. Importantly, the study-time task allows investigation of several learning phenomena in greater detail than has been possible by using only naturally occurring knowledge structures such as routine scripts in texts, or person stereotypic descriptions. Our method enables researchers to vary the training history and structure of a given category (schema) as well as its relation to other categories. The method also allows investigators to observe and track trial-by-trial the effects of a schema on selective attention, encoding, and memory performance. In this manner, the task allows tests of theories of how categories are discovered and learned in the first place, especially in unsupervised learning tasks. Such virtues should provide helpful tools for further research into major properties of cognitive learning.

References

Anderberg, M. R. (1973). In *Cluster analysis for applications*. New York: Academic Press.

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge, & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension*. Hillsdale, NJ: Erlbaum.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413-451.

Anderson, J. R. (1991a). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.

Anderson, J. R. (1991b). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, *12*, 587-625.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*, 177-220.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, *2*, 331-350.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Carroll, J. D. (1976). Spatial, non-spatial, and hybrid models for scaling. *Psychometrika*, *41*, 439-463.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Clapper, J. P., & Bower, G. H. (1991). Learning and apply category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory. Vol. 27*. New York: Academic Press.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 443-460.

Davis, B. R. (1985). An associative hierarchical self-organizing system. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*, 570-579.

deGroot, A. D. (1965). *Thought and choice in chess*. Mouton: The Hague.

deGroot, A. D. (1966). Perception and memory verus thought. In B. Kleinmuntz (Ed.), *Problem-solving*. New York: Wiley.

Estes, W. K. (1994). In *Classification and cognition*. New York: Oxford University Press.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 556-571.

Everitt, F. (1980). In *Cluster analysis*. London: Heinemann.

Feigenbaum, E. A. (1963). A simulation of verbal learning behavior. In E. A. Feigenbaum, & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139-172.

Fisher, D., & Langley, P. (1990). The structure and formation of natural categories. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory.* (pp. 241-284). San Diego, CA: Academic Press.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgements. *Cognitive Psychology*, *16*, 341-370.

M., Gluck,, & J., Corter,. (1985). Information, uncertainty, and the utility of categories. In *Proceedings of the seventh annual conference of the Cognitive Science Society* (pp. 283-287). Hillsdale, NJ: Erlbaum Publishers.

Graesser, A. C., Woll, S. B., Kowalski, D. J., & Smith, D. A. (1980). Memory for typical and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 503-513.

Hanson, S. J., & Bauer, M. (1989). Conceptual clustering, categorization, and polymorphy. *Machine Learning*, *3*, 343-372.

Heit, E. (1993). Modeling the effects of expectations on recognition memory. *Psychological Science*, *4*, 244-252.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory. Vol. 18.* New York: Academic Press.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241-254.

Kolodner, J. L. (1983). Reconstructive memory: A computer model. *Cognitive Science*, *7*, 281-328.

Kosslyn, S. M., & Pomerantz, J. R. (1977). Imagery, propositions, and the form of internal representations. *Cognitive Psychology*, *9*, 52-76.

Lebowitz, M. (1982). Correcting erroneous generalizations. *Cognition and Brain Theory*, *5*, 367-281.

Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning*, *2*, 103-138.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*, 159-188.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in research and theory. Vol. 24.* New York: Academic Press.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review*, *85*, 207-238.

Miller, G. A. (1969). A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, *6*, 169-191.

Millward, R. B. (1971). Theoretical and experimental approaches to human learning. In J. W. Kling, & L. A. Riggs (Eds.), *Experimental psychology. Third edition* (pp. 905-1017). New York: Holt, Rinehart & Winston.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw Hill.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 104-114.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. M. (1988). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87-108.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Postman, L. (1971). Transfer, interference, and forgetting. In J. W. Kling, & L. A. Riggs (Eds.), *Experimental psychology* (3rd ed.) (pp. 1019-1132). New York: Holt, Rinehart & Winston.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97*, 285-308.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Richman, H. B. (1991). Discrimination net models of concept formation. In D. H. Fisher, M. J. Pazzani, & P. Langley (Eds.), *Concept formation: Knowledge and experience in unsupervised learning*. San Mateo, CA: Morgan Kaufman.

Richman, H. B., & Simon, H. A. (1989). Context effects in letter perception: Comparison of two theories. *Psychological Review, 96*, 417-432.

Rosch, E. (1975). Cognitive representation of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Advances in cross cultural psychology*. New York: Academic Press.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1* (pp. 45-77). Cambridge, Mass.: MIT Press.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the aquisition of knowledge.*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Sattah, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*, 319-345.

Schank, R. C. (1982). *Dynamic memory.* Cambridge, UK: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Hillsdale, N. J.: Lawrence Erlbaum Associates.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting and clustering. *Science, 210*, 290-298.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*, 125-140.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review, 96*, 58-83.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 30*, 379-393.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford: Blackwell.

Footnotes

This research was supported by a grant, 91-0144, from the Air Force Office of Scientific Research, and a grant, MH-47575, from the National Institute of Mental Health.

Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, CA 94305.

Figure Captions

Figure 1. Sample stimulus sets illustrating how categories may be defined in terms of correlated attribute values. There are 8 attributes (columns) with binary values (1 or 2). Each row represents an instance. In panels a and c, the 8 instances in the left block reflect one category (called "A") and those in the right block reflect a second category ("B").

Figure 2. Computer display as it appeared during the three phases of Experiments 1 and 2. Panela illustrates an instance display with all features masked except for the bark-color attribute. Panel 6 illustrates the 4-alternative recognition memory test. Panel c illustrates feedback of the correct answer following a subject's choice on the memory test.

Figure 3. Recognition-memory accuracy and study-time data from Experiment 1. Trials are shown in their original order in this figure; the functions are disconnected to indicate where the A- and B-blocks are separated in the Blocked condition, and where the test block begins in all three conditions. Study-time in seconds is abbreviated as "ST" on the vertical axis of this and the following figures.

Figure 4. Recognition-memory accuracy and study-time data from Experiment 2. Trials are shown in their original order, and the functions are separated as in Figure 3.

Figure 5. Recognition-memory accuracy and study-time data from Experiment 3. Pretraining trials are shown in their original order and separated from the test trials which follow. The test trials are separated by category in both conditions (the A-trials are shown before the B-trials, although the trials were mixed.

Figure 6. Recognition-memory accuracy and study-time data from Experiment 4. Trials are shown in their original order, and the pretraining trials are separated from the test trials.

Figure 7. Recognition-memory accuracy and study-time from Experiment 5. Trials are shown in their original order, with pretraining and test blocks separated.

Figure 1

a)

Attribute

1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	2
1	1	1	1	1	1	2	1
1	1	1	1	1	1	2	2
1	1	1	1	1	2	1	1
1	1	1	1	1	2	1	2
1	1	1	1	1	2	2	1
1	1	1	1	1	2	2	2

Attribute

1	2	3	4	5	6	7	8
2	2	2	2	2	1	1	1
2	2	2	2	2	1	1	2
2	2	2	2	2	1	2	1
2	2	2	2	2	1	2	2
2	2	2	2	2	2	1	1
2	2	2	2	2	2	1	2
2	2	2	2	2	2	2	1
2	2	2	2	2	2	2	2

Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

b)

Attribute

1	2	3	4	5	6	7	8
1	2	2	2	2	2	1	1
1	2	2	1	2	1	2	2
2	1	2	2	2	2	1	2
1	1	2	2	1	1	2	1
2	2	2	2	1	1	1	2
2	1	1	2	1	2	2	1

Attribute

1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	2
2	2	2	1	2	2	2	2
2	1	1	1	2	2	2	1
2	2	1	1	2	2	2	1
1	1	1	1	2	1	1	2
1	2	1	2	1	2	1	2

No categories defined

c)

Attribute

1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1
1	1	1	1	1	3	1	1
1	1	1	1	1	1	1	1
1	3	1	1	1	1	1	3
1	1	1	3	1	1	1	1
1	1	3	1	3	1	1	1
1	1	1	1	1	1	3	1

Attribute

1	2	3	4	5	6	7	8
2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	4
4	2	2	4	2	2	2	2
2	2	4	2	2	2	2	2
2	2	2	2	2	4	4	2
2	2	2	2	4	2	2	2
2	4	2	2	2	2	2	2
2	4	2	2	2	2	2	4

Category "A" : 1 1 1 1 1 1 1 1
Category "B" : 2 2 2 2 2 2 2 2

Figure 2

```
a.      Aralia

        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        dark grey bark
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

        Press INS or DEL to see next item
```

```
b.      Aralia

        1. deep brown bark
        2. dark grey bark
        3. mossy green bark
        4. light tan bark


        ******************************
        * Enter a number from 1 to 4 *
        ******************************
```

```
c.      Aralia

        1. deep brown bark
    --> 2. dark grey bark
        3. mossy green bark
        4. light tan bark

           INCORRECT

    Arrow indicates correct choice

        Press RETURN to go on
```

Figure 3a

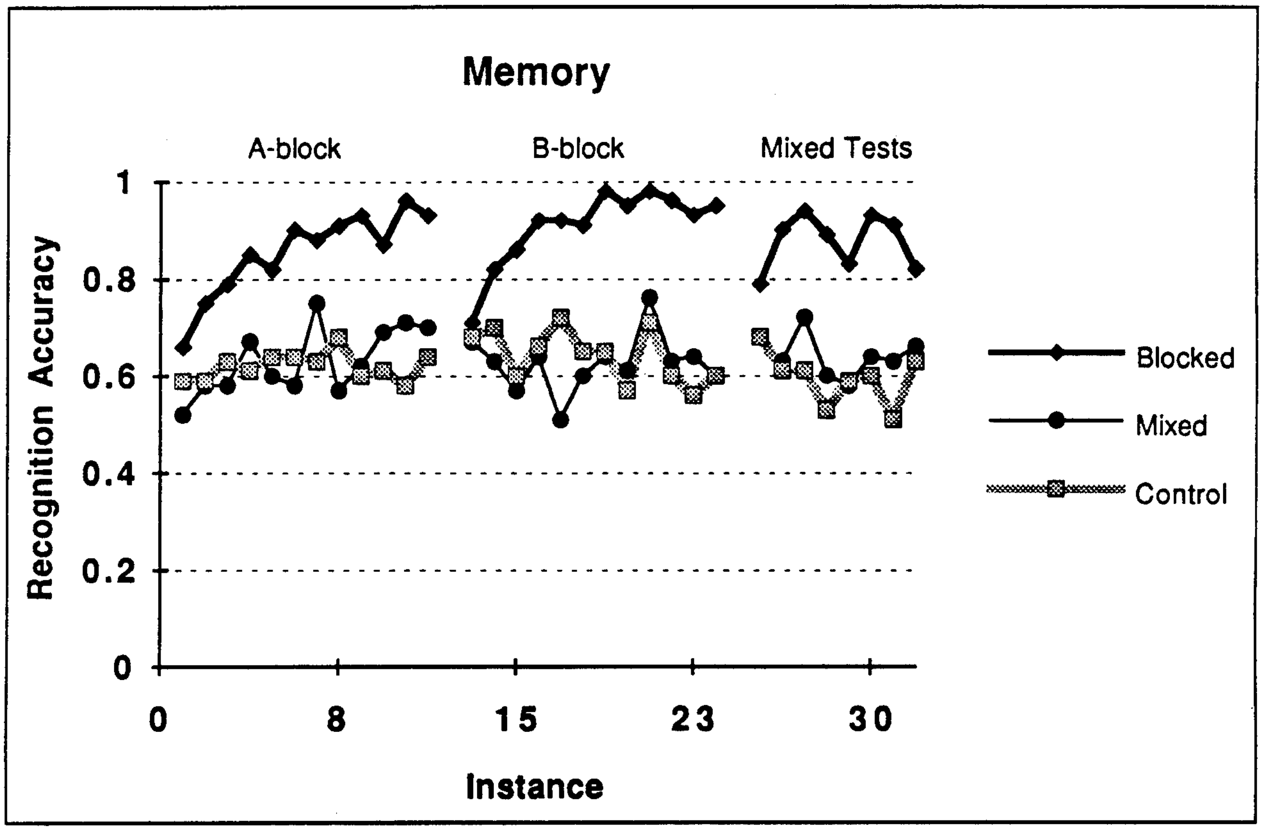

Memory
A-block
B-block
Mixed Tests
Recognition Accuracy
1
0.8
0.6
0.4
0.2
0
0
8
15
23
30
Instance
Blocked
Mixed
Control

Figure 3b

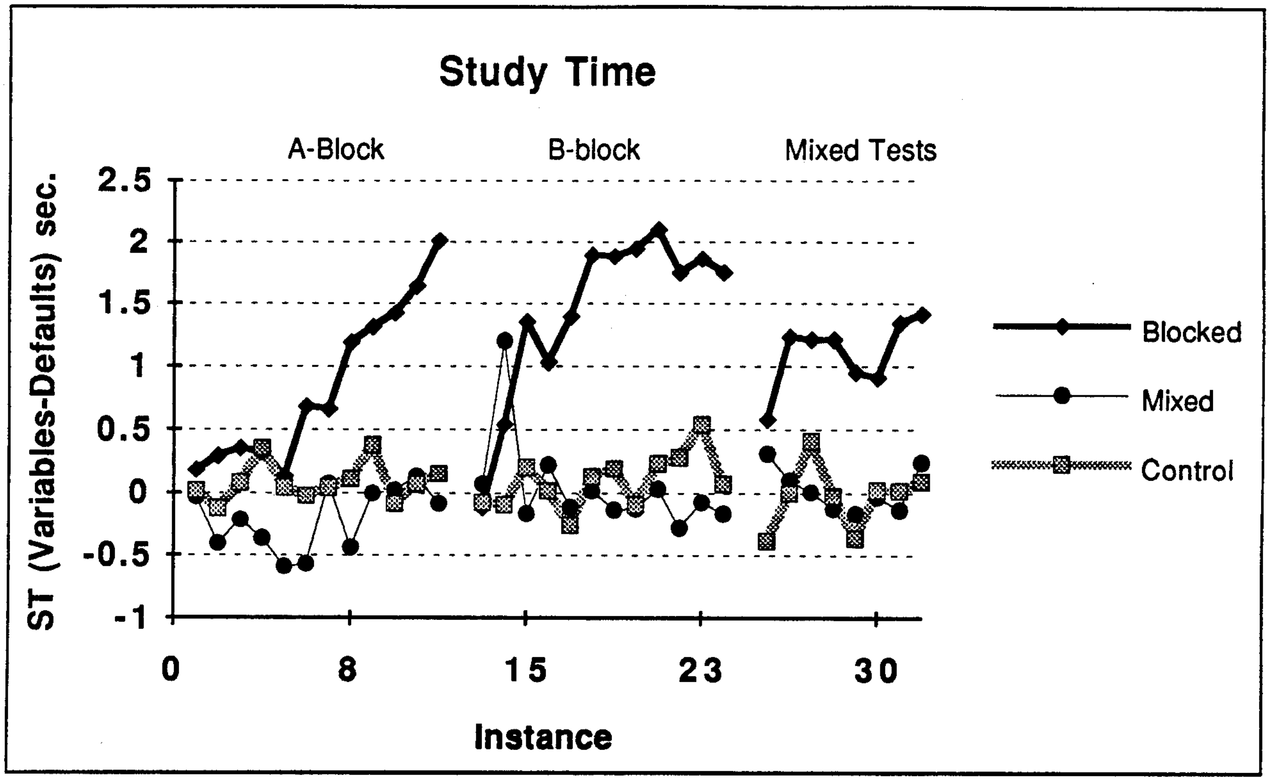

Study Time
A-Block
B-block
Mixed Tests
ST (Variables-Defaults) sec.
2.5
2
1.5
1
0.5
0
-0.5
-1
0
8
15
23
30
Instance
Blocked
Mixed
Control

Figure 4a

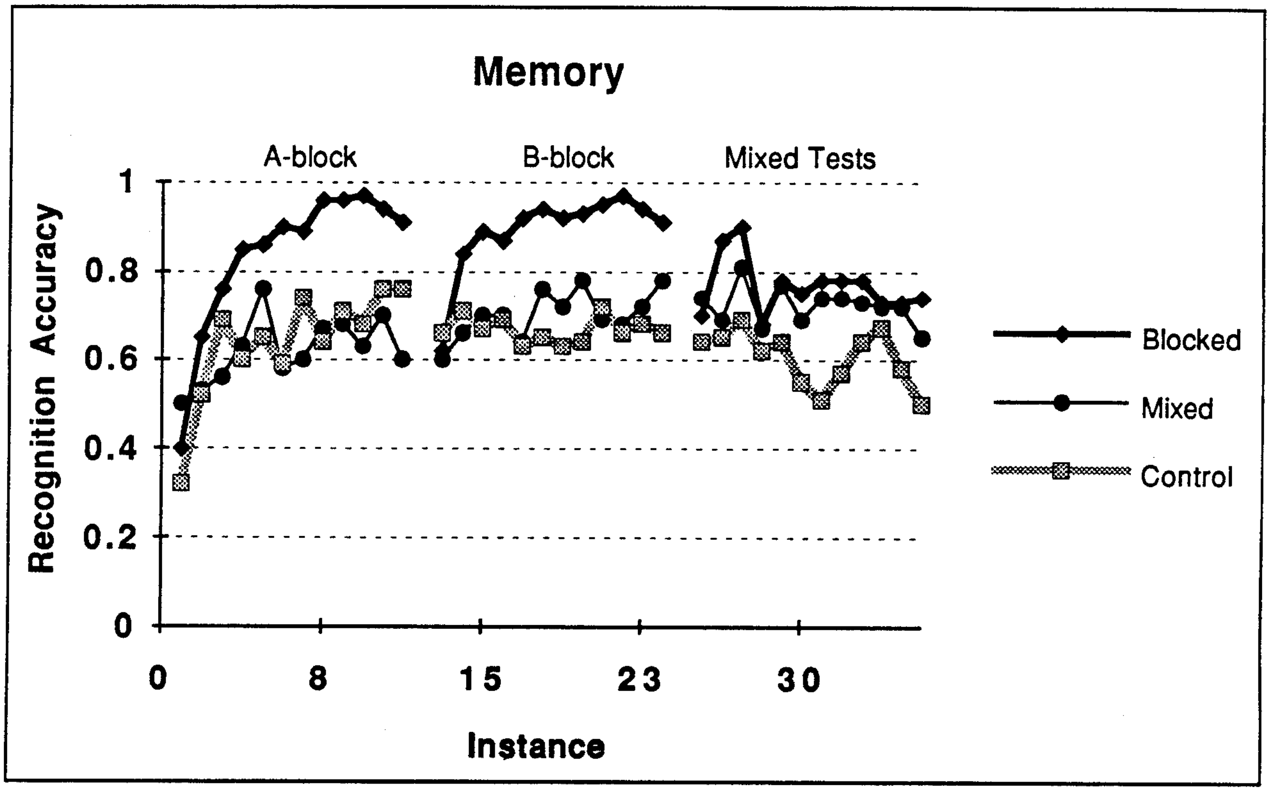

Memory
A-block
B-block
Mixed Tests
Recognition Accuracy
1
0.8
0.6
0.4
0.2
0
0
8
15
23
30
Instance
Blocked
Mixed
Control

Figure 4b

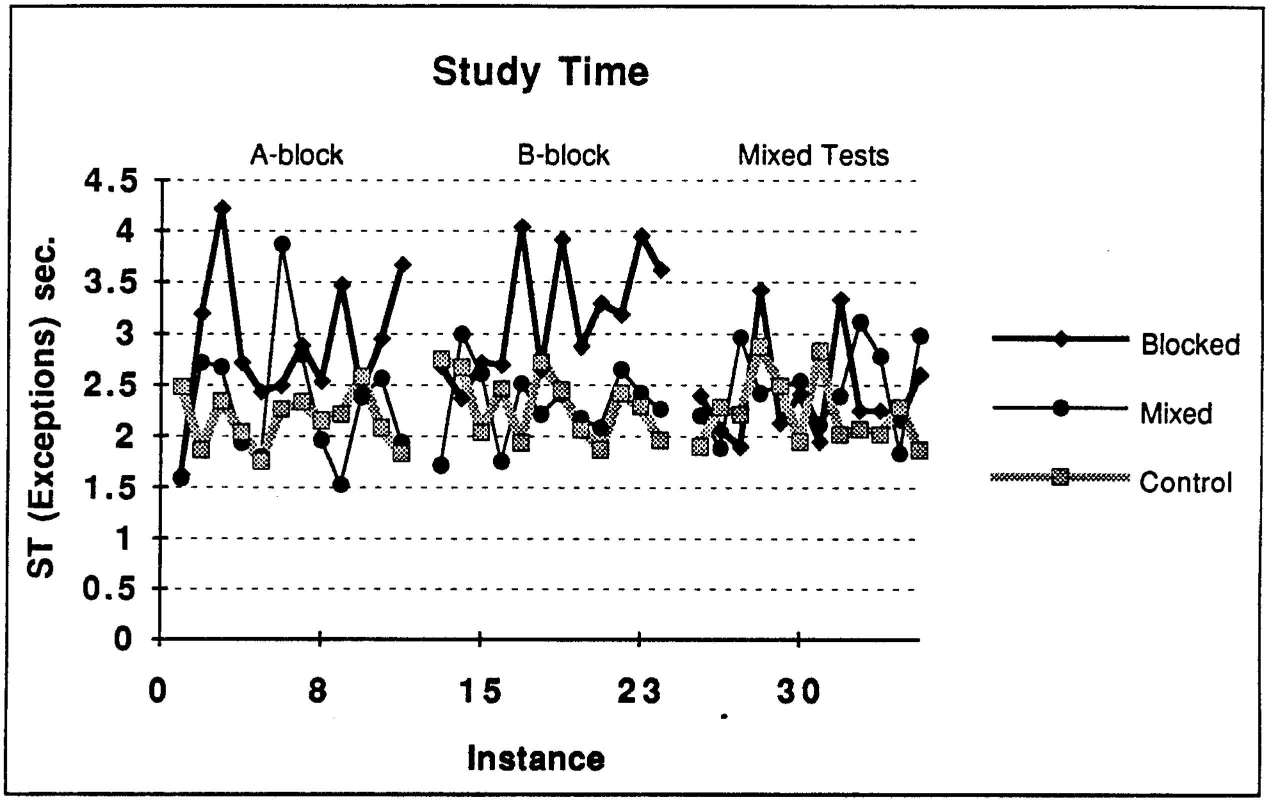

Study Time
A-block
B-block
Mixed Tests
ST (Exceptions) sec.
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
8
15
23
30
Instance
Blocked
Mixed
Control

Figure 5a

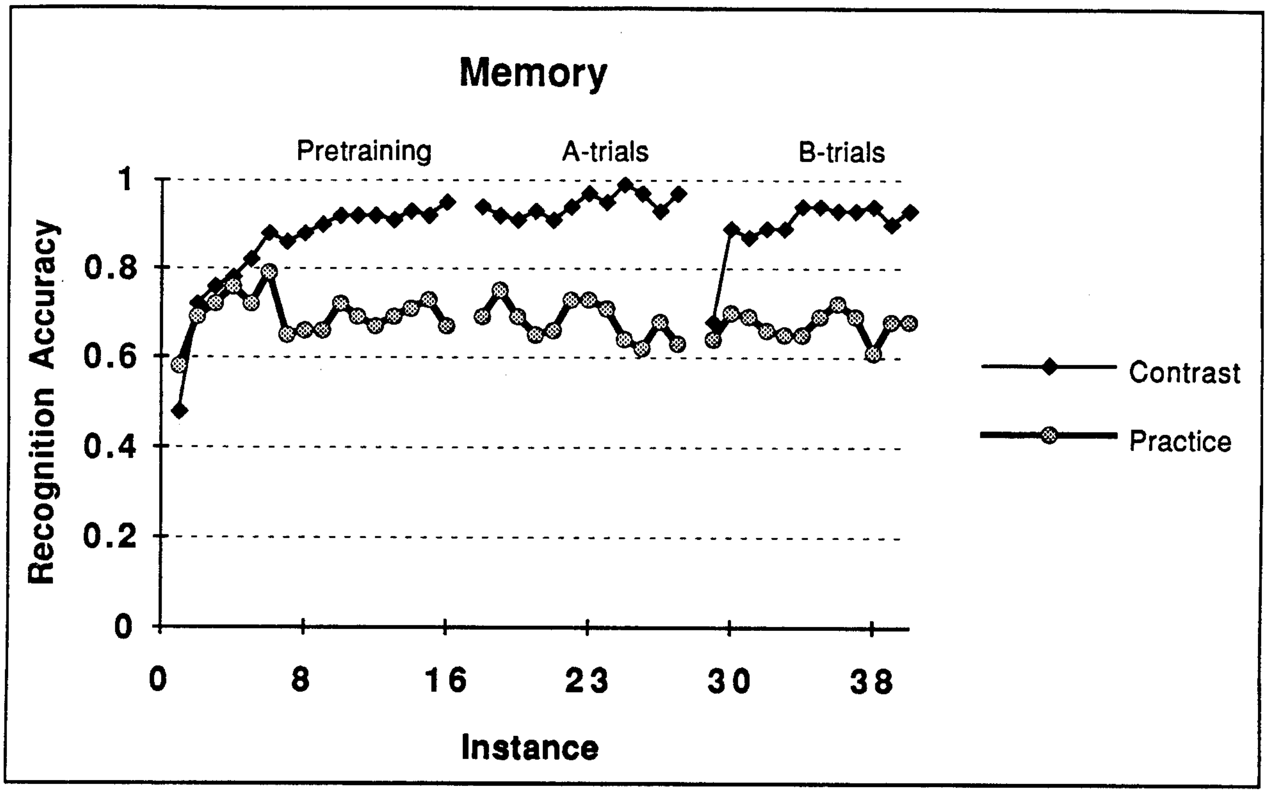

Memory
Pretraining
A-trials
B-trials
1
0.8
0.6
0.4
0.2
0
Recognition Accuracy
0
8
16
23
30
38
Instance
Contrast
Practice

Figure 5b

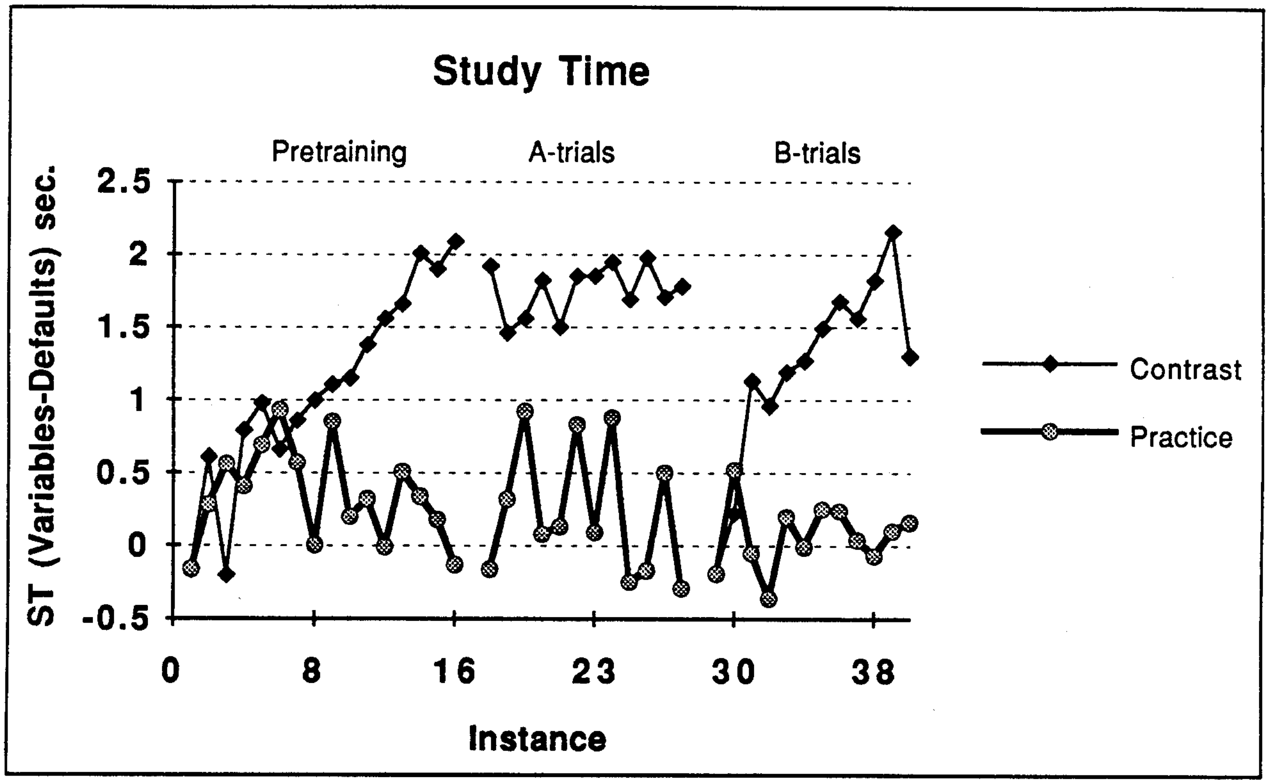

Study Time
Pretraining
A-trials
B-trials
ST (Variables-Defaults) sec.
2.5
2
1.5
1
0.5
0
-0.5
0
8
16
23
30
38
Instance
Contrast
Practice

Figure 6a

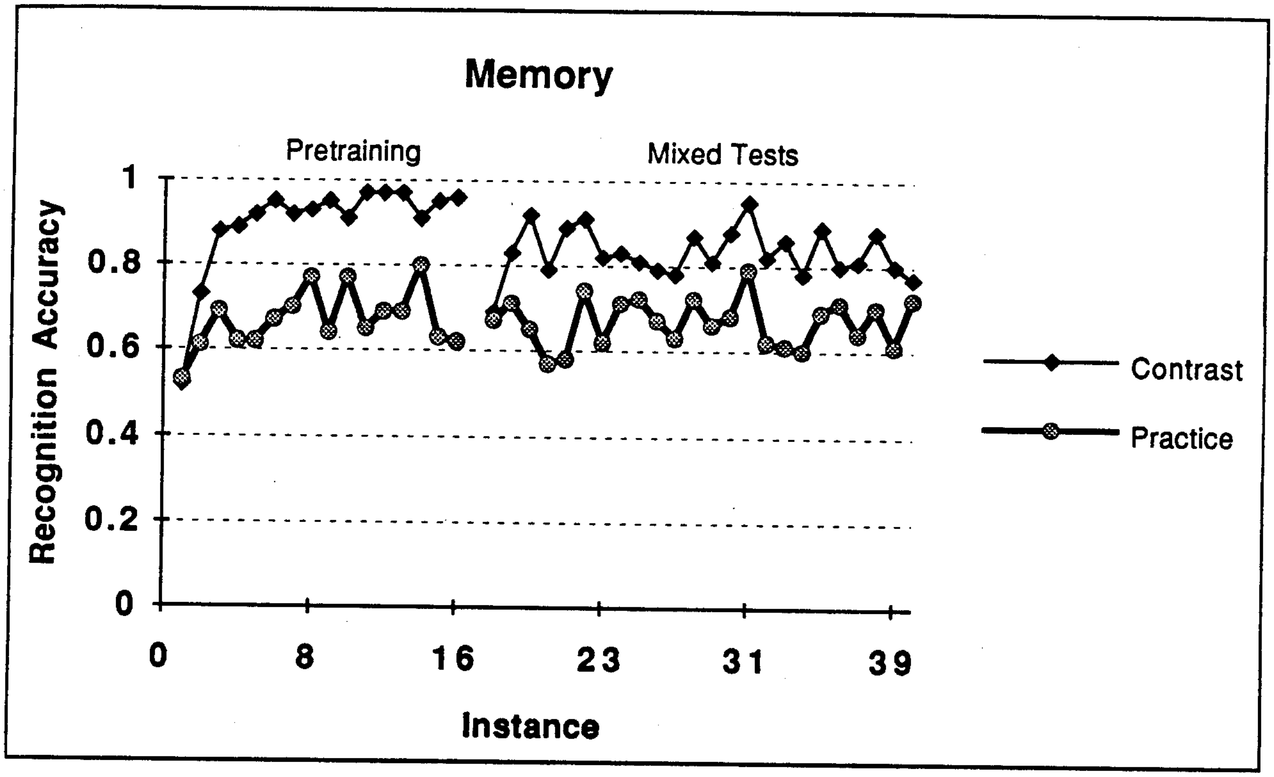

Memory
Pretraining
Mixed Tests
Recognition Accuracy
1
0.8
0.6
0.4
0.2
0
0
8
16
23
31
39
Instance
Contrast
Practice

Figure 6b

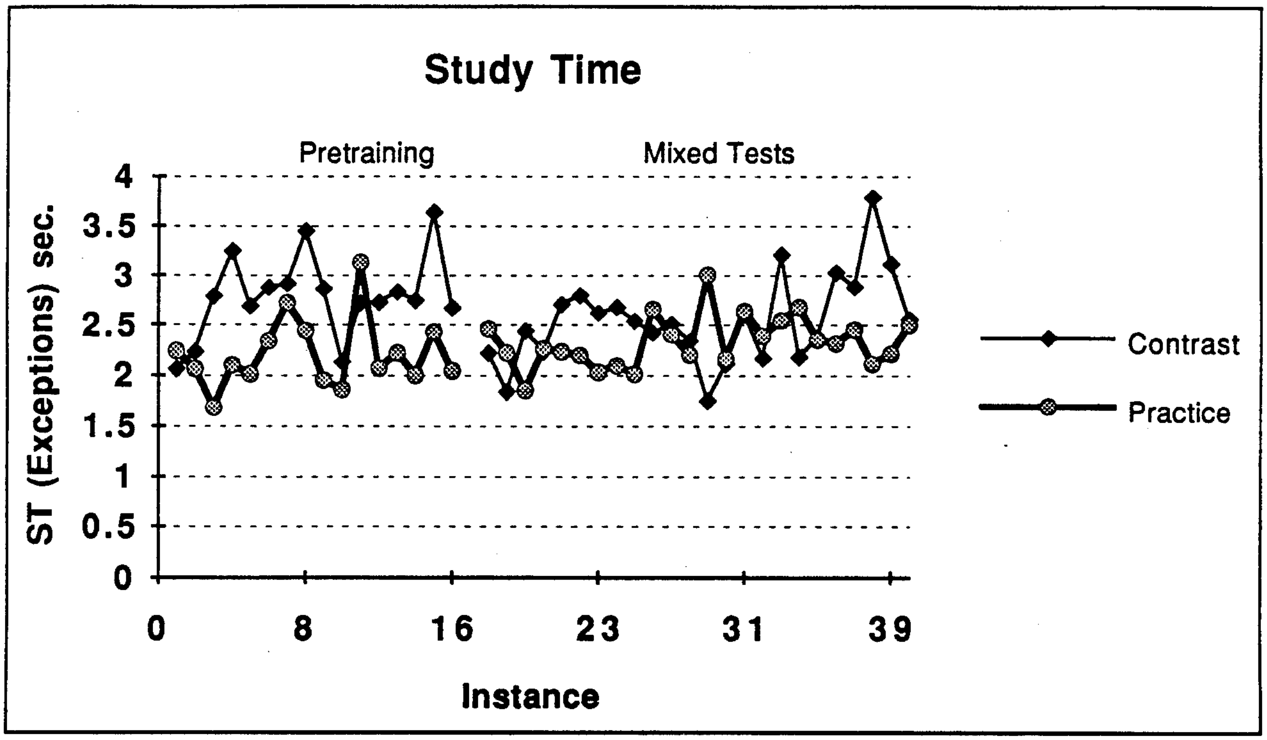

Study Time
Pretraining
Mixed Tests
ST (Exceptions) sec.
4
3.5
3
2.5
2
1.5
1
0.5
0
0
8
16
23
31
39
Instance
Contrast
Practice

Figure 7a

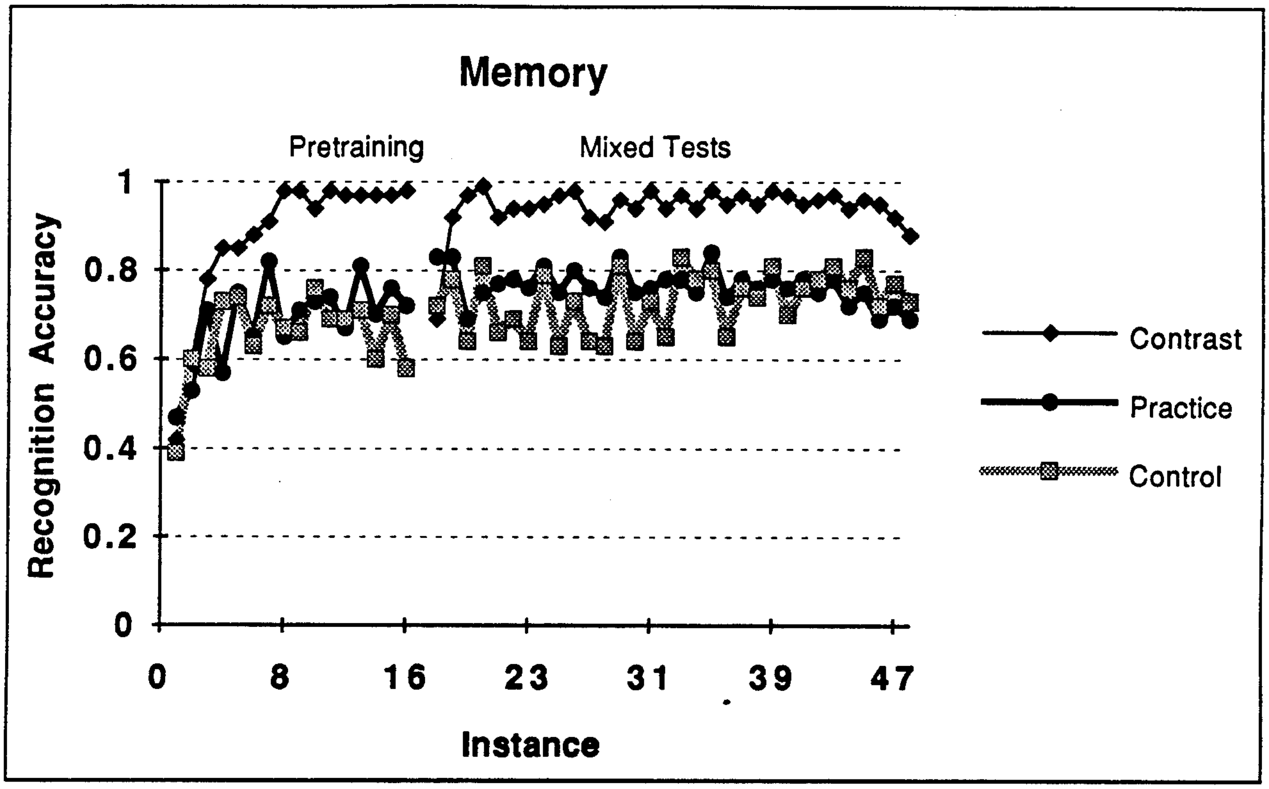
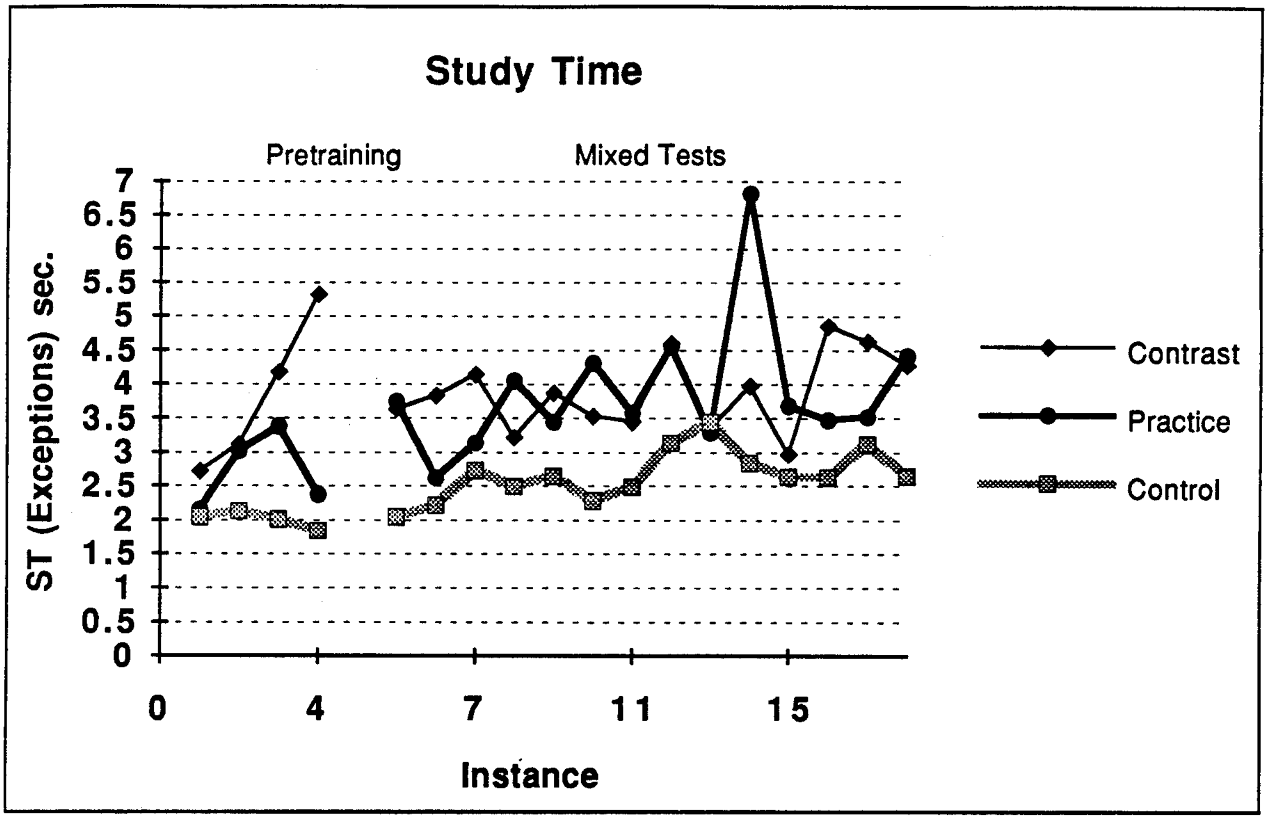

Memory
Pretraining
Mixed Tests
Recognition Accuracy
1
0.8
0.6
0.4
0.2
0
0
8
16
23
31
39
47
Instance
Contrast
Practice
Control

Figure 7b


Study Time
Pretraining
Mixed Tests
ST (Exceptions) sec.
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
4
7
11
15
Instance
Contrast
Practice
Control

Category Invention and Transfer of Learning in Unsupervised Tasks

John P. Clapper and Gordon H. Bower

Stanford University

ADDRESS ALL CORRESPONDENCE TO:

Gordon H. Bower
Department of Psychology, Bldg. 420
Stanford University
Stanford, CA 94305
(415) 387-5544

Running Head: UNSUPERVISED LEARNING

Abstract

Three experiments investigated the principles by which categories are discovered and prior category knowledge is used to facilitate learning categories without feedback (unsupervised learning). These experiments provided evidence that categories are discovered primarily through their contrast with previous categories, and that subjects create new categories as needed based on contrasts rather than by accumulating evidence about patterns of correlated features over successive training instances. When a new category shared default features with a previous category, subjects appeared to learn new norms mainly for the features on which the two categories differed, transferring norms for shared features to the new category rather than relearning them. The data also suggest that the context in which a category is learned and the type of stimulus materials used (verbal or pictorial) can influence the later stability or retention of separate categories. The results are discussed in terms of the constraints they place upon acceptable models of human category learning.

Introduction

The ability to group objects and events into discrete categories, and to learn generalized descriptions of these categories, is one of the fundamental components of human intelligence. Two general experimental situations have been used to study category learning. In supervised learning tasks, subjects are asked to sort a series of training instances into predefined categories, and are given regular feedback on the accuracy of their classifications. Such feedback is absent from unsupervised learning tasks, in which subjects must invent their own categories based on any informative regularities or patterns they detect within the stimulus set.

While an extensive body of research has accumulated on supervised learning (see, e.g., Bruner, Goodnow, & Austin, 1956; Homa, 1984), unsupervised learning has received much less attention within experimental psychology. Few established procedures or reliable measures have been developed for investigating unsupervised learning. In this article, we employ procedures developed by Clapper and Bower (1991, 1994b) to investigate the processes by which people discover and learn about categories in unsupervised tasks. These procedures provide reliable indices of subjects' unsupervised category learning as they examine a series of training instances to perform tasks which, on the surface, appear unrelated to categorization.

Defining unsupervised learning

To evaluate unsupervised category learning in experimental tasks, it is first necessary to define what kinds of patterns or structures within a stimulus set are considered to give rise to distinct categories. We will be concerned here with categories defined in terms of correlated stimulus features, i.e., properties that consistently occur together over time and separate presentation episodes. To make this notion more precise, let us begin by characterizing stimulus sets in terms of abstract dimensions or *attributes*, each of which can assume one or another of a range of concrete *values*. Specific attribute values of a training instance will also be referred to as *features* of that instance. To illustrate, artificial stimulus sets for laboratory experiments are often generated from a few general attributes such as shape (with values triangle or square), size (large or small), and color (red or green), so that individual stimuli within the sets are distinguished by specific values on each attribute, e.g., a large red triangle vs. a small green square. Although natural stimuli are generally more complex than such artificial stimulus sets, in theory natural stimulus sets could also be characterized in terms of fixed sets of attributes.

When a domain is characterized by given attributes, categorical structure may be defined in terms of correlations or patterns of co-occurrences among the values of different attributes. For example, the stimulus set illustrated in Figure 1a displays perfectly correlated values on the first 5 of the 8 attributes listed, but not the last 3a (the order of listing is arbitrary). These correlational patterns define two distinct categories within the stimulus set, each category corresponding to a particular cluster of co-occurring attribute values. By contrast, in Figure 1b all the attributes of the stimuli vary independently, providing no basis for partitioning the stimulus set into meaningful categories. Within a correlated stimulus set, the consistently co-occurring values that give rise to a given category (e.g., the first 5 in Fig. 1a) will be referred to as the *default* attribute values of that category. Uncorrelated attributes (e.g., the last 3 in Fig. 1a) will be referred to as *variable* attributes, and specific values of these attributes will be called variable values or simply variables.

Within unsupervised tasks, category learning can be defined in terms of subjects' sensitivity to the correlational structure of a stimulus set, i.e., responding discriminately to the underlying patterns that give rise to categorical distinctions. Such learning is demonstrated by showing that the correlational patterns influence subjects' performance on some task -- for example, by showing that memory is improved when a stimulus set contains correlational patterns, or that memory for correlated values exceeds that for uncorrelated values.

Measures of unsupervised learning

Clapper and Bower (1991, 1994a, 1994b) developed several indirect indices of unsupervised learning of such correlation-based categories. The primary effect of learning such correlational patterns is an increased ability to predict the features of an instance given partial information about that instance, such as its category membership. Given that subjects can remember the category to which an instance belonged, they can reconstruct the values of all its correlated attributes on the basis of general norms acquired from previous instances. Thus, subjects show improved memory for default values within a correlated set, compared to the features of otherwise equivalent instances from uncorrelated sets (Clapper and Bower, 1991, 1994b).

By definition, default values are predictable within a given category, and since defaults are shared by all or most instances within that category, they provide no basis for distinguishing among different instances. Different instances within a category are distinguished in terms of their variable, non-default, features. For example, in Fig. 1a, the patterns in the first and last rows are distinguished from one another only by their values on the last three (variable) attributes. Therefore, when memorizing a given instance subjects need only remember what category it belongs to, plus the variable values which are not predictable from this category membership. The subjects would not need to encode default values as features of the specific instance, because these values could be inferred from general norms previously acquired about its category. Much evidence has shown that, as people learn stimulus materials based on familiar natural categories, they tend to encode these materials by remembering the category (script or schema) which they instantiate, plus information that could not be inferred from this categorization. (see, e.g., Bower, Black & Turner, 1979; Clapper & Bower, 1991, 1994a; Graesser, Woll, Kowalski & Smith, 1980; Schank, 1982; Schank & Abelson, 1977). This is referred to as the "schema-plus-corrections" (S+C) encoding model.

In previous research, Clapper and Bower (1994b) showed that subjects who learned correlational categories conformed to the S+C encoding strategy, and that use of this strategy could even be taken as an index of category learning. Thus, subjects asked to memorize a series of training instances with correlated attribute values spent significantly more time studying unpredictable variables than predictable defaults of the instances. This tendency increased with the degree of category learning, and provided a reliable basis for comparing learning in different training conditions. Due to this increased attention to variable features while studying the instances, category learning resulted in improved memory not only for predictable defaults, but also for the values of the unpredictable attributes (compared to a control condition in which the stimuli were composed of uncorrelated attributes).

A related index of learning, employed by Clapper and Bower (1991, 1994a), asked subjects to write down a list of those attribute values of each training instance that distinguished it from similar instances they had seen, while omitting features that would not provide such discriminating information. They were asked to record only those features that would enable them to recognize this instance in a later multiple-choice test. Since variables distinguish different instances within a category whereas defaults do not, as subjects learned the correlated defaults within the presented stimulus sets, they increased their listings of variables while decreasing that of defaults. The difference in frequency of listing defaults versus variables provided an index of category learning similar to that provided by differences in study times in the instance memorization task.

Methods of unsupervised learning

Given a definition of categories within an unsupervised task, as well as procedures for measuring subjects' learning of them, we have tried to discriminate among different processes by which such learning might occur. Clapper and Bower (1994a, 1994b) distinguished two general processes by which correlational patterns could be acquired in the absence of predefined categories or categorization-related feedback. [1] These differ primarily in whether categories are viewed as arising through the accumulation of correlational data over a series of instances, or as arising through the discrete hypothesizing of distinct categories triggered by detecting explicit mismatchs and contrasts between instances from different categories.

Clapper and Bower (1994a, 1994b) referred to the first general approach as *autocorrelation*, because in such models a person's knowledge of correlational structure is represented directly, in the form of a matrix of correlational associations or as a collection of correlational rules. As each new training instance is encountered, the associations (or correlational rules) relating the attribute values of that instance to one another are strengthened. As additional instances are presented from a given category, associations between consistently correlated values increase in strength, while associations between values that do not occur consistently together will be weaker. With training, a learner would acquire the ability to predict the presence of default values given other default values, and to distinguish correlated defaults from uncorrelated variables. For example, in the categories of Figure 1a, a value of 1 or 2 in any of the first five relevant attributes perfectly predicts a corresponding value of 1 or 2 in the remaining five attributes. In the autocorrelation approach, categories need not be represented explicitly because all category information would be captured within the network of explicit correlational associations. Several models of this type are represented in the literature, e.g., single-layer connectionist models such as those of J. A. Anderson and others (see, e.g., Anderson, 1977; Anderson, Silverstein, Ritz & Jones, 1977; Rumelhart, Hinton, & McClelland, 1986) or rule-learning models such as that of Billman and Heit (1988) and Davis (1985).

A second hypothesis regarding people's learning of correlational structure assumes that people can invent (hypothesize) new categories as needed to capture whatever patterns or regularities they notice in a given stimulus set. We refer to this method as *category invention* (Clapper and Bower, 1994a, 1994b). In the absence of any direct autocorrelation, a learner must use some other basis for deciding when to create new categories. One alternative to autocorrelation is to invent new categories whenever existing categories fail to accommodate a novel or surprising training instance. To illustrate, imagine that subjects are shown the eight instances of category A from Figure 1a, followed by two instances of category B. We assume that a new category is created at the start of training to describe the first instance of A, and that further A instances are then assimilated to this category. When the first instance of category B is presented, it will violate the default values of the first 5 attributes that had been previously acquired for category A. We assume that if the learner's subjective impression of surprise or mismatch exceeds some internal criterion (e.g., a majority of defaults are violated), then learners will create a separate category to cover this B stimulus, and proceed to assimilate further B instances to this new category and further A instances to the original A category. By assigning instances that exemplify different correlational patterns to separate categories, and computing general norms (central tendencies such as averages or probability distributions of values) within each category, a learner could internalize much of the same information as contained in a direct correlational record but without needing to keep track of associations among all possible feature combinations. This mismatch or contrast heuristic for creating new categories is similar to the "failure driven learning" of Schank (1982), and to the "novelty detectors" used to reclassify instances in some multi-layered connectionist models (e.g., Carpenter & Grossberg, 1987).

In category invention, a new category is created because some of the features of a novel instance contradict strong default expectations of the most similar existing category (called the reference category). Importantly, the contrast between the first new (B) instance and the previous (A) category depends on the strength or confidence of the default norms subjects have acquired about category A. If subjects have seen only a few instances of category A and so have relatively low confidence regarding the default norms for this category, then the first instance of category B will not result in the failure of strong expectations, and hence it may not trigger the creation of a new category.[2] If not, then instances of both A and B categories will be included in the original category; thus subjects would fail to properly conditionalize the correlational structure of the stimulus set (i.e., all the attributes would be encoded as independent variables). The stronger and more definite the default expectations learned about category A prior to the presentation of category B, the greater the subjective impression of contrast or mismatch between the two patterns, and the greater the probability that a separate category will be created to describe the B instances.

According to category invention, discovering and distinguishing categories based on different correlational patterns is an inherently comparative, sequential process. Each category is learned by contrast with categories that have been acquired previously. When presentations of two different categories are intermixed from the start of training and must be learned concurrently, subjects may perceive no strong contrast between them and thus the instances may be lumped together into a single category, obscuring the correlational structure within each category. Thus, two-category learning will be reduced by any manipulation which changes a sequential learning problem, in which category B can be learned by its contrast to strong defaults of a prior-learned A category, to a concurrent problem, in which A and B instances are intermixed from the start of training so that both sets of defaults must be learned concurrently. This outcome should hold true even when the manipulation results in a larger number of instances from a given category being shown in the concurrent condition than in the sequential condition. To illustrate, a condition in which ten instances of category A are presented prior to a mixed sequence of As and Bs may be compared to another condition in which a random 5 of the first 10 instances of category A are replaced by instances of category B. Even though the number of B instances is larger in the second condition, the structure of the learning series has changed from sequential to concurrent presentations of the two categories; this change should decrease the probability of creating separate categories and thus reduce average learning of category B.

Autocorrelation models are relatively unaffected by subjective contrast and rely exclusively upon experiences of feature co-occurrences (accumulation of correlational strengths over successive instances of a category) to acquire correlational patterns. A learning process that depends exclusively upon autocorrelation expects that, all else being equal, learning of a given correlational pattern should increase monotonically with exposure to instances of that pattern. Therefore, autocorrelation expects learning of category B to be higher in the concurrent condition described above than in the sequential condition; according to an autocorrelation approach, the only important difference between conditions is the different numbers of instances shown from each category.

Clapper and Bower (1994a, 1994b) reported several experiments in which reducing the number of instances presented from a given category increased learning of that category, apparently due to heightened contrast of the second category with a well-learned, earlier category. This pattern of results, in which perceived contrast had a larger impact on learning than the number of instances shown from a given category, was referred to as a *contrast effect,* and was interpreted as evidence for category invention rather than autocorrelation in unsupervised learning.

The Clapper & Bower experiments provided evidence for explicit category invention in several situations. Contrast effects were obtained in both attribute-listing and instance memorization tasks, with both pictorial stimuli and verbal stimuli, and for categories characterized by correlational patterns in which defaults were present with 100% reliability as well as those in which the correlations were less than perfect (Clapper & Bower, 1994a, 1994b). That evidence was limited, however, because the situations investigated involved only two categories that were always distinguished by contrasting values on the same attributes. Possibly, 2-category learning is a special case in which contrast plays an unusually large role, and perhaps when more categories are being learned autocorrelation would play a larger role relative to between-category contrasts. These possibilities are tested in the experiments described below, all of which use more than 2 categories and in which some of the categories contrast on all attributes while others within the same set have partially overlapping defaults. If significant contrast effects were observed in such situations, we would have convincing evidence of the generality of contrast effects and category invention in unsupervised learning.

Transfer of learning in category invention

Category invention implies that new norms must be hypothesized and stored in memory to represent a new category. This description raises the question of how new categories are represented in relation to old categories, i.e., what information is added to the learner's existing conceptual knowledge to capture the novel properties of the new category?

In the present context, the new (triggering) stimulus is assumed to contain a contrasting set of default attribute values within the same set of attributes as the reference category. The minimum requirement for capturing such discrepancies is that new norms be created for each attribute of the triggering stimulus that violates norms of the reference category. However, the learner need not create new norms for aspects of the triggering instance that are shared with the reference category; these shared values would constitute the basic set of attributes by which both subcategories would be described. Such information could be omitted from the norms of the new category so long as the learner encoded this category explicitly as a set of modifications to the reference category. In this case, features which the new category shares with the reference category could be retrieved indirectly from the norms of the reference category.

This minimal representation strategy can be viewed as a normative or rational rule for category invention, because it maximizes efficiency in terms of encoding resources and memory organization. A new category is acquired by making the fewest possible modifications to the learner's existing knowledge base required to accurately represent how the new category differs from previously acquired ones. Granting that new learning (memorizing new cognitive structures and their elements) requires cognitive resources (i.e., time or encoding capacity), then category invention in humans should approximate this normative model to some degree. The logical alternative, an exhaustive full-copy recording of all defaults of the new category, seems implausible given the complexity of natural stimuli and categories and the difficulty of many real-world learning problems.

This approach to creating new categories is consistent with the standard S+C model, described above, of how people encode specific instances of familiar categories (see, e.g., Clapper & Bower, 1991, Graesser et al., 1980; Schank and Abelson, 1977). When subjects encounter the first few instances of a new category, they should allocate attentional resources to attributes in a manner predicted by the S+C model. That is, they should attend most to those features of the new instances that contradict the norms of the reference category, after glancing briefly at (then passing over) defaults from the reference category that are shared by instances of the new category. Due to this lack of attention at encoding, such shared defaults should tend to have lower weight or salience in the memory representations of these new instances and the category norms based on them. Thus, the S+C encoding process tends to produce economical category invention as a by-product of its economical strategy for encoding individual instances.

For convenience in describing such related categories, we will refer to the shared defaults as *superordinate* defaults, because by virtue of being shared by more than one category these features define an implicit superordinate category which contains the subcategories possessing those defaults. Default values which distinguish each subcategory within this broad superordinate category will be referred to as *subordinate* defaults. To illustrate, consider the following four categories, depicted in terms of the same abstract numerical codes used to describe the categories in Figure 1: A1 = 111111XX, A2 = 111222XX, B1 = 333333XX, and B2 = 333444XX. In this example, the first two categories (A1 and A2) share three out of six default values, as does the second pair (B1 and B2). The stimulus set can be partitioned into two superordinate categories (A vs. B) distinguished by two contrasting values on the first three attributes. In turn, each superordinate category may be partitioned into two subordinate categories (A1 vs. A2, and B1 vs. B2), distinguished by four contrasting values on the second three attributes. The instances within a subcategory are distinguished by values on the last two variable attributes, the Xs. Thus, categories that overlap in this manner comprise a two-level default hierarchy; such hierarchies play a central role in many theories of knowledge representation (see, e.g., Clapper & Bower, 1991; Collins & Quillian, 1969; Collins & Loftus, 1975; Holland, Holyoak, Nisbett & Thagard, 1986; Schank, 1982 Kolodner, 1984).

We expected that when a new subcategory was introduced following an earlier subcategory sharing the same superordinate default values, subjects would maintain the same level of attention and memory for these shared defaults as they had shown on the last instance of the prior reference category; but we also expected subjects would show sharply increased attention to, and decreased memory for, new subordinate defaults for a few trials. Such heightened attention to subordinate attributes was expected to diminish as additional instances of the new subcategory were encountered, due to strengthening of the new subordinate default values to the level of the shared defaults. In these initial studies our main concern was to provide basic demonstrations of default transfer effects, i.e., to design laboratory situations in which category invention could be shown to conform to the basic S+C framework. Later research may then determine the specific factors that influence transfer performance and that could serve as a basis for discriminating among alternative models of transfer.

Experiment 1

The goals of the present experiment were (1) to corroborate earlier evidence that people learn categories in unsupervised tasks primarily through category invention rather than autocorrelation, and (2) to investigate how people transfer default norms from previously acquired categories to partially overlapping new categories.

To provide evidence for learning by category invention, we compared learning in two different sequencing conditions. In the Contrast condition, 3 categories were introduced into the training sequence one at a time, so that subjects could learn the defaults of a given category very well before seeing any instances of a new category. Such a presentation sequence should facilitate category invention by maximizing the subjective contrast between the first instance of each new category and the default norms acquired about previous categories. The second (Mixed) condition presented instances of the different categories in a randomly intermixed sequence from the start of training. In this condition, the collections of default values characterizing the different categories would have to be acquired concurrently. The category invention hypothesis predicts that subjects encountering several related categories in such a mixed sequence without supervision may lump them together into a single overgeneralized category, and thus fail to learn the conditionalized correlational structure of the stimulus set. Thus, subjects should show poorer learning of the categories in this Mixed condition than in the Contrast condition, even though a larger number of instances from a given category had been shown in the Mixed condition. Such a contrast effect cannot be accommodated within a learning process based purely upon autocorrelation.

A second goal of the present experiment was to demonstrate the transfer of all relevant norms from previously-learned categories to new subcategories. Of the three categories shown, two (A1 and A2) shared the same values on several defaults (superordinate defaults) and differed along several others (subordinate defaults). The third category (B) differed from the two on both super- and subordinate defaults. Category A1 was presented first in the Contrast condition, followed by categories A2 and B. Since category A2 had the same superordinate defaults as the prior category A1, subjects were expected to show significant positive transfer in acquiring these superordinate defaults, compared to the new subordinate defaults. No such difference should be observed for category B, which shared neither superordinate nor subordinate defaults with the previous A categories.

This experiment employed the instance memorization task introduced in Clapper and Bower (1994a), which provides two distinct indices of category learning on each trial (viz., study-time and recognition memory). If observed, the predicted pattern of results would provide evidence for the generality of the category invention process, and indicate that this process takes advantage of prior knowledge to economize on new learning and attentional resources.

Method

Subjects

The subjects were 38 undergraduate students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement.

Procedure

Subjects were tested in groups of 10 to 15 for a single 90 min session. Each subject was seated in front of an individual microcomputer terminal, which administered all aspects of the experiment. After subjects read the instructions presented on the computer screen and signed a form indicating their informed consent to participate, the main portion of the experiment began.

Each trial consisted of a study phase followed by a test phase. At the beginning of the study phase, a list display was presented in the middle of the CRT screen. At the top of the list was the name of a fictitious tree instances (these were arbitrarily selected Latin names from a plant identification guide), below which appeared a vertical listing of twelve verbal feature descriptors, one per row. At the start of the trial, each descriptor was masked by a row of X's (see Figure 2a). Starting from a random attribute (row) showing in the list, subjects studied the feature descriptors by pressing a designated "line up" or "line down" key which removed the X's on the line below or above the current line and allowed them to examine its item (attribute value). The exposed attribute value was covered up again as soon as the inspection pointer was moved to a new line (attribute). This procedure permitted subjects to allocate their study time among the features any way they wished, within the constraint that the total study time equaled 36 secs. The computer recorded the total amount of time spent looking at each attribute.

Insert Figure 2 about here

Each item was a verbal description of a specific value of a particular stimulus attribute. For example, the attribute "color of bark" had four alternative values, such as "dark grey" and "mossy green". The attributes were presented in the same serial order (screen locations) on each trial, although different values of a particular attribute could occur on successive trials. That is, the color of the tree's bark might always appear as the fifth spatial position, its growth rate in the seventh position, and so on. In this manner, subjects could learn the locations of the default and variable attributes.

After a study interval of 36 sec, the list disappeared and the test phase began. During this phase, subjects were tested on their memory for the values of all twelve attributes of the just-presented instance. The test items were presented one at a time in a multiple-choice format (see Figure 2b). The name of the most recent instance appeared at the top of the multiple-choice display with four alternative answers below. These alternatives were always different values of the attribute being tested, e.g., four different habitat preferences or growth rates. Subjects tried to remember which of these values had occurred in the just-studied instance and typed in the number corresponding to that choice on their computer keyboard. Following this response, the computer displayed either a "correct" or an "incorrect" prompt under the test display, which remained on the screen. If the response was incorrect, the correct choice was indicated by an arrow in the display (see Figure 3c). The subject then pressed a "Next" key.

After answering all twelve test questions about a given instance, subjects received summary feedback for the trial. The percentage of items answered correctly on that trial was displayed, and below this the average percentage correct pooled over all test trials completed up to that point. If the trial score was higher than the cumulative score, the message "Good job! You beat your overall score!" appeared on the screen; if not, the message "Try to beat your overall score next trial" was displayed. If the subject answered all the test questions correctly on a given trial, the message "Good job! Your score was perfect!" was displayed.

The twelve attributes were tested in a different random order on each trial, and the order in which values were listed in the multiple-choice test display was also randomized separately on each trial. The experiment consisted of a total of 52 such instance study-test trials. Following this, subjects read a debriefing sheet that informed them of the purpose and methods of the experiment.

Materials and Design

The training instances were verbal descriptions of fictitious trees, presented in a list format. The instances were characterized in terms of twelve substitutive attributes, each with four possible values, defining a stimulus set of 4^{12} possible instances. In 3 of these attributes, only 2 values ever occurred in the training instances; in 7 other attributes, 3 of the 4 values were shown, and all 4 values were shown in the remaining 2 attributes. All 4 values of each attribute were shown during the multiple-choice testing.

The stimuli in this experiment were divided into 3 different categories. These categories were defined in terms of patterns of correlated attribute values, and are referred to as categories A1, A2, and B. Representing these categories by numerical codes similar to those displayed in Figure 1, category A1 was 1111111111xx, category A2 was 1112222222xx, and category B was 3333333333xx. Particular serial positions in these codes correspond to attributes of the stimuli, and the numbers appearing in those positions indicate default values of the corresponding attributes. The x's in the last 2 positions indicate variable attributes, which varied independently over 4 different values within all three categories. Note that categories A1 and A2 shared the same default values on the first 3 of their attributes (superordinate defaults) but had different default values on 7 other attributes (subordinate defaults). Category B had a different set of both super- and subordinate default values than either of the A categories.

Subjects were randomly assigned to two different conditions, which differed in the number of instances shown from each category and the order in which they occurred. In the *Contrast* condition, instances of category A1 were shown for the first 12 trials. Over the next 18 trials, 12 instances of category A2 and 6 instances of category A1 were presented in random order. Following this, 12 instances of category B, 6 instances of category A2 and 4 instances of category A1 were shown over the last 22 trials, again in random order. Note that in this sequencing, subjects saw only instances of A1 during the first block of 12 trials, instances of both A1 and A2 in the second block of 18 trials, and saw instances of all three categories in the final block of 22 trials. There was nothing in the procedure to cause subjects to notice this separation between blocks, except for the instances of new categories not shown previously.

In the *Mixed* condition, all three categories were shown from the start of training. The first block of 12 trials showed 4 instances of each category. The second block (of 18 trials) contained 12 instances of A2, 3 instances of A1, and 3 instances of B. Note that this second block had the same number of instances of A2 as the second block of the Contrast condition, but in the Contrast condition the 6 remaining trials all showed instances of A1 instead of showing both A1 and B instances. The third block in the Mixed condition was the same as the corresponding block of the Contrast condition, showing 4 instances of A1, 6 of A2, and 12 instances of B. This will be referred to as the test block in both conditions, and the two preceding blocks will be referred to as training blocks. The order of instances within each block was randomized.

The instances within each trial block were constructed so that each value of the variable attributes occurred an approximately equal number of times, and so that specific combinations of variable values did not recur within the same block. Over the experiment as a whole, each of the 16 possible combinations of values from the two variable attributes was used an approximately equal number of times. These steps were taken to ensure that subjects did not encounter consistent correlations between variable values that might cause them to form spurious subcategories or correlational rules.

Balancing

The stimuli for all the subjects in a given condition were generated by the testing program from the same input file, which contained coded specifications for generating the instances presented on each trial. Within a given block, the stimuli were presented in a different random order for each subject. The correspondence between serial positions in the codes and the order in which an attribute was listed in the training instances was randomized for each subject. These random assignments were undertaken to balance out any idiosyncratic effects due to particular attributes, values, or combinations of values on the experimental data.

Results and Discussion

The data collected in this experiment were the time subjects spent studying each attribute value of an instance (study times or STs), and their accuracy in remembering each attribute value during the multiple-choice testing phase of each trial. The study time data and recognition memory data are displayed in Figure 3. Due to the random sampling procedure used, data was collected from a total of 21 subjects in the Contrast condition and 17 subjects in the Mixed condition. Data analysis in this and later experiments involved large numbers of cooperative t-tests of significance. Rather than tediously reporting all of these many pairwise comparisons, we will adopt throughout a $p < .05$ criterion for statistical significance and simply state which comparisons are significant by that criterion. Readers interested in actual t's and dfs may consult the authors.

Insert Figure 3 about here

In the Contrast condition, subjects learned the categories sequentially, acquiring strong norms for each category before instances of new categories were introduced into the training sequence. As in previous research on sequential category learning (e.g., Clapper & Bower, 1994b), each category was learned rapidly as it was encountered. Recognition memory (averaged over attributes) for Contrast subjects increased significantly from 0.60 on the first trial to an overall average of 0.89. A second index of learning was computed by averaging STs over super- vs. subordinate default attributes and then subtracting this average from the ST for variable attributes. This difference index was highly significant averaged over trials in the Contrast condition.

By comparison, little evidence for category learning was obtained in the Mixed condition. Averaged over trials and categories, the ST difference index yielded no significant evidence of learning ($p > .25$). In addition, recognition memory showed less improvement over trials than in the Contrast condition. Memory averaged 0.71 over trials, significantly less than the average of 0.89 from the Contrast condition.

Contrast effects and category invention

Fewer instances of categories A2 and B were presented in the Contrast condition than in the Mixed condition, but the order of instances in the Contrast condition was arranged so that the categories could be learned sequentially, whereas the default norms of all three categories had to be acquired concurrently in the Mixed condition. If subjects relied mainly on contrast and explicit category invention to distinguish the categories, then the Mixed condition should show poorer learning than the Contrast condition.

The data confirmed the expected contrast effects, providing strong evidence that learning by category invention is not restricted to the 2-category situations investigated by Clapper and Bower, (1994a, 1994b). Category A2 occurred 12 times during the second block in both Contrast and Mixed conditions; however, 4 instances of this category were presented during the first block of the Mixed condition, whereas none occurred prior to the second block in the Contrast condition. Although more instances of A2 were seen by subjects in the Mixed condition, instance memory was higher in the Contrast condition (0.91 vs. 0.79). Subjects in the Contrast condition also spent significantly more of their study period attending to variable attributes of the instances of subcategory A2 than did subjects in the Mixed condition.

Following this second block, an additional 6 instances of category A2 were shown during the third block in both conditions. Mixed condition subjects increased their attention to variable attributes of A2 during this third block; perhaps they were finally beginning to learn the category defaults of A2 after seeing 16 to 20 instances. This change eliminated the previous difference in STs between the Mixed and Contrast conditions for A2 during this final block. However, overall memory for instances of A2 remained significantly higher in the Contrast condition than in the Mixed condition.

Thus, the superior learning of category A2 observed in the Contrast condition during in the second block was sustained into the third block.

The pattern of results was similar for category B. Subjects in the Contrast condition saw no instances of category B prior to the third block, whereas subjects in the Mixed condition saw 7 instances of category B prior to this block. Nonetheless, memory for category B instances during the third block was significantly better in the Contrast condition. The Contrast condition also showed marginally higher STs for variable attributes than did the Mixed condition ($p < .10$).

Transfer of learning

When creating new categories, a rational learning process should avoid unnecessary duplication by transferring the attribute structure and confirmed defaults of the reference category to the new category. This process characterizes the S+C model of category invention. In the present experiment, this process implies that subjects should learn category A2 by modifying their representation of category A1, retaining the superordinate defaults shared by the two categories while creating new norms to describe the new subordinate defaults of A2. Thus, when first encountering A2 instances, subjects should increase their attention (ST) to subordinate defaults, but not to superordinate defaults and not to variable attributes. Despite this increased attention, memory for the new subordinate defaults should be reduced for the initial instances of A2, compared to previous trials in which these attributes had well-learned A1 default values. Memory for the A2 subordinate defaults should increase thereafter as subsequent instances of A2 are encountered and the new default norms are strengthened. In contrast to this pattern for subordinate defaults of A2, memory for superordinate defaults shared by categories A1 and A2 should not be affected by the switch from category A1 to A2, but should remain at the same high level as in the preceding A1 block. Since memory for variables depends on the degree of overall default learning, it should decrease during the early instances of A2, but then improve as the new A2 defaults are learned.

The S+C model predicts a different pattern of results when category B is introduced to the training sequence. Since both the super- and subordinate defaults of category B differ from those of the two previous categories, learners would need to create new norms for both sets of features as they acquired the B category. Thus, STs for both sets of features should increase when the first instance of category B is presented, and both sets should show decreasing STs as subsequent B instances are encountered and the new defaults strengthened. In parallel, memory performance should show a drop-off for both types of default attributes and variables when the first instance of category B is presented. However, memory for the defaults and the variables should increase as the B defaults are learned over the next few trials.

Study-time Data

The ST data in Figure 3a were largely consistent with these predictions. For category A2, STs to subordinate defaults increased significantly (0.72 sec) on the first instance compared to the preceding A1 trial, while STs to superordinate defaults decreased nonsignificantly (0.31 sec, $p > .10$). The large difference in STs between super- and subordinate defaults on this trial significantly exceeded the corresponding difference from the previous A1 trial. STs to variables also decreased sharply on this trial, because subjects allocated a larger share of the fixed study period to encoding the new subordinate defaults, reducing the share left over for variables.

Following this first trial, STs to subordinate defaults decreased significantly over the next few instances of category A2, reflecting their increasing strength in subjects' category norms. Superordinate STs increased significantly over the same interval, while the difference in STs between these two attribute types decreased, but remained significant over the block as a whole. The increase in superordinate STs presumably occurred because the proportion of the fixed study period available to study these values would have increased as STs to subordinate defaults decreased over the first few instances of A2.

A different pattern of ST results was observed when subjects encountered the first instance of category B. Compared to the preceding trial, STs to both superordinate and subordinate defaults increased on the first category B trial, although only the increase in superordinate STs attained the 5 percent level of statistical significance. STs for both superordinate and subordinate defaults decreased slightly over the first 3 instances of category B, in contrast to the differential pattern shown over the first few instances of A2.

To summarize, the ST data showed a pattern of transfer predicted by the S+C model of category invention. Most importantly, there was a significant difference in transfer of super- vs. subordinate defaults from A1 to A2, whereas no such difference was observed at the origination of category B. The patterns of changing STs for super- and subordinate defaults were also different over the early trials of A2 compared to the corresponding B trials.

Instance Memory Data

The memory data are consistent with the evidence for transfer from the ST data, but suggested that the transfer of superordinate defaults from A1 to A2 was incomplete or imperfect. Memory for superordinate defaults decreased somewhat on the first A2 trial, compared to the previous A1 trial, and then increased to its previous level over the next few instances. This initial reduction in memory for superordinates was not predicted by the S+C transfer model.

Memory for subordinate defaults was slightly lower than that for superordinates on the first A2 trial, but this difference was not significant ($p < .25$). Memory for subordinate defaults remained slightly depressed for the rest of the A2 block, significantly so compared to the superordinates and the prior A1 block.

A somewhat different pattern of results was observed in memory for category B. For the first instance of this category, memory for subordinate defaults remained at the same level as on the previous A2 trial. However, memory for superordinates decreased significantly on this trial, and then recovered to its previous levels on the following trial. The sharp initial reduction in superordinate memory might have occurred because subjects had learned to ignore superordinate defaults over the preceding A1 and A2 instances, and were thus caught off-guard when these defaults were finally violated in the first instance of category B.

In sum, the ST data showed patterns of transfer consistent with the S+C model of category invention. However, the memory data failed to confirm the model's prediction that memory for superordinate defaults would remain at previous levels in the first instances of A2. This lack of transfer in the memory measure might have been caused by limitations in our subjects' memorization abilities. If subjects were unable to remember all the attributes in which the first instance of A2 differed from their prior norms for A1, this confusion could have disrupted their transfer performance on the memory tests. Given uncertainty about which values were changed and which had been left the same, subjects might have made errors in verifying both super- and subordinate values. Since this source of uncertainty was not present during encoding, when all features of an instance were present for subjects to examine, it would not have affected observed transfer on the ST measure. These issues will be discussed in greater detail below in the General Discussion.

Experiment 2

Category invention implies that correlational patterns should be acquired more easily in sequential than in concurrent presentation conditions. Two forms of sequential presentation may be distinguished: (1) a pure blocking arrangement in which categories are presented separately in unmixed blocks or series of instances, e.g., several instances of category A followed by several instances of category B; and (2) a partially blocked arrangement, illustrated by the Contrast condition of Experiment 1, in which categories are learned one at a time as in the fully blocked sequence, but in which prior categories continue to be shown as each new category is being learned.

Previous research on unsupervised learning (Clapper & Bower, 1994a, 1994b) has shown high levels of learning in both blocked and contrast sequences when subjects learned two contrasting categories. This result indicates that little retroactive interference (RI) was exerted by the later category upon subjects' retention of the earlier category in these 2-category experiments, and that little forgetting of the earlier category occurred due to simple delay. However, it seems likely that interference among the categories would be more severe when three categories are shown, as in Experiment 1, and when partial overlap among the defaults of these categories makes them more easily confused. If this were the case, then the contrast sequence should show more stable category learning than the blocked sequence, because subjects in the contrast sequence continue to see instances of previously acquired categories, thus maintaining their learning of these categories as they acquire new ones.

Besides maintaining the learning of prior categories, the contrast sequence might facilitate explicit comparisons between different categories as they are acquired. Each time a training instance is classified in such a sequence, subjects receive practice in distinguishing new from old categories. In a fully blocked sequence, on the other hand, instances of previous categories are no longer shown once a new category is introduced to the training sequence. As a result, the norms of these previous categories would become less available from memory over subsequent trials of a new-category block. In this situation, instances could be categorized more or less by default after the first few trials of a new block, and subjects thus would receive less practice in explicitly distinguishing between the different categories as they classified each training instance.

This difference in the trial-by-trial need for norm availability could affect what subjects learn about the new category and the representation of this learning in memory. Specifically, subjects in the Contrast condition might tend to form representations of each category that contain explicit cues or mnemonics to help them tell the categories apart. Such distinguishing features would likely be less prominent in subjects' representations of categories learned in the comparative isolation of a fully blocked training sequence. As a result, subjects might experience difficulty in transferring norms of categories learned in isolation to a mixed sequence in which these categories must be distinguished from other, closely related categories. When such categories are shown in a mixed sequence, subjects might be more likely to confuse the different correlational patterns, mixing up the defaults of different categories and losing confidence in their norms for each category.

The design of Experiment 2 allowed us to compare details of unsupervised learning in such fully blocked vs. contrast sequences. The same mixed test block was shown at the end of training in both conditions, allowing us to compare final levels of learning and assess retention of categories learned earlier in both conditions. The experiment also allowed us to separate the effects of simple forgetting from the effects of the learning context. The last category presented in the Blocked condition should not be forgotten prior to the mixed test block; thus, if this category shows weaker learning during the test block than the corresponding category from the Contrast condition, this difference could not be due to greater forgetting in the blocked sequence. Rather, it would most probably be related to the differing contexts in which the categories are acquired in the two sequencing conditions.

The stimulus set used in Experiment 2 had the same categories and correlational structure as that used in Experiment 1, thus providing an opportunity to replicate those transfer effects and confirm the S+C model of category invention. Such transfer in learning new categories was expected in both conditions of the present experiment.

Method

Subjects

The subjects were 28 students of San Jose State University participating in partial fulfillment of their Introductory Psychology course requirement.

Procedure

The experimental procedure was identical in most respects to that of Experiment 1. Subjects were tested in groups of 10 to 15 for a single session lasting approximately 90 minutes. Each subject was individually seated at a computer terminal in a single large testing room. The entire experiment, consisting of 54 trials plus instructions and debriefing, was administered by computer. As before, each trial consisted of a 36 sec study phase followed by a multiple-choice testing phase in which subjects had to choose which of the four possible values of each attribute had been presented in the last instance. The computer recorded how long subjects spent looking at each attribute as well as their recognition memory accuracy for each attribute during the test phase.

Materials and Design

The learning instances and categories were the same as in Experiment 1. That is, subjects learned the three categories A1 = 1111111111xx, category A2 = 1112222222xx, and category B = 333333333xx.

Subjects were randomly assigned to two different conditions. The *Contrast* condition was similar to that of Experiment 1. Only instances of category A1 were presented during the first block of 12 trials in this condition. Following this, 10 instances of category A1 and 12 instances of category A2 were presented in the second block in random order. During the third block, subjects saw 6 instances each of categories A1 and A2, and 8 instances of category B, presented in random order. As in Experiment 1, there was no break or separation between blocks, except that a new category was presented at the start of each block.

In the second, *Blocked* condition, categories were initially separated in the training sequence prior to being presented together in a mixed sequence near the end of training. The first block was the same as that of the Contrast condition, i.e., 12 instances of category A1. This was followed by a second block in which 12 instances of category A2 were presented, and then by a third block containing 10 instances of category B. The fourth block contained the same instances as the third block of the Contrast condition, i.e., 6 instances of category A1, 6 instances of A2, and 8 instances of category B. As in Experiment 1, this final block will be referred to as the test block in both conditions, and all the preceding blocks will be referred to as training blocks.

Balancing the Design

As in Experiment 1, stimuli were generated from coded specifications from a computer file. The order of instances within a block of trials (described above) was shown to each subject in a different random order, but the blocks themselves were presented in the same order to all subjects within a given condition. The assignment of attributes (serial positions) in these instance codes to concrete stimulus attributes such as "bark color" or "leaf shape" was randomized separately for each subject. Care was taken to ensure that each value of the variable attributes occurred with approximately equal frequency within each block and over the experiment as a whole. The same was true for each of the 16 possible combinations of the two variable values. The order in which attributes were listed during the study phase was also randomized separately for each subject, and remained constant for that individual throughout the experiment. The order in which attributes were queried during the multiple-choice testing, as well as the order in which the alternative values were listed, varied randomly over trials for each subject.

Results and Discussion

The same recognition memory and ST data were collected in this experiment as in Experiment 1 and are displayed in Figure 4.

Insert Figure 4 about here

The Contrast condition showed much the same pattern of learning as the corresponding condition from Experiment 1. Each of the three categories were learned rapidly according to both the ST and memory indicators, and this learning was retained over subsequent blocks (the second and third blocks for category A1 and the third block for category A2). The Blocked condition also showed strong learning of all three categories during the training phase. However, both the memory and ST indices indicated that performance based on learning of all three categories in the Blocked condition decreased during the final test block, when instance of the three categories were presented in a mixed sequence.

During the training phase, according to the ST measure, learning appeared a bit higher in the Blocked condition than in the corresponding trials of the Contrast condition whereas the memory data appeared to show a bit more learning in the Contrast condition. However, neither of these differences were statistically significant (p*p* < .10).

Although learning in the two conditions was roughly equal during the training phase, performance in the Blocked condition appeared lower during the test block, when categories previously shown separately were presented together in a mixed sequence. As a result, default learning appeared stronger in the Contrast condition than in the Blocked condition during this test block. All three categories were remembered significantly better in the Contrast condition. Moreover, STs to variable attributes were also higher in this condition.

Importantly, learning was higher in the Contrast condition even though more instances had been shown from a given category in the Blocked condition, i.e., from categories A2 and B. A particularly interesting result was the significant difference in final learning of category B. Although subjects in the Blocked condition had been shown 10 successive instances of category B immediately prior to the test block whereas subjects in the Contrast condition had seen no previous instances of this category, learning was significantly lower in the Blocked than the Contrast condition.

Besides revealing poorer learning during the test block, the Blocked condition also performed unequally with the three categories in a manner related to their order of acquisition. The average variable ST during the test block was 3.67 seconds for instances from the first category (A1), 3.28 seconds for instances of the second category (A2), and 3.83 for instances of the third category (B). The test for a quadratic trend over these three categories was marginally significant, ($p < .10$). This ST data thus showed a U-shaped pattern of category learning that resembled the familiar "serial position effect" observed in verbal learning (Murdock, 1962).

The memory data appeared to show a slight serial position effect in its pattern of means during the test block (.91 for category A1, 0.86 for A2, and 0.88 for B), but the quadratic test was nonsignificant in this data ($p > .15$). The corresponding ST and memory data from the Contrast condition showed no evidence of such a serial position effect, with no significant differences between the category means of either variable STs or overall memory.

Is it forgetting?

One hypothesis to explain better final performance in the Contrast condition is that subjects might forget earlier categories as they acquired later ones in the Blocked condition, whereas the continued practice on earlier categories prevented such forgetting in the Contrast condition. While this forgetting explanation explains the apparent loss of A1 and A2 learning between training and test phases of the Blocked condition, it fails to explain the lowered performance for category B during the test block, since this category was learned immediately prior to the test block and no significant forgetting could have occurred over this interval. The fact that category A1 showed equal performance to category B during the final test block casts further doubt on the forgetting explanation, since that would expect better performance on the more recent category.

These arguments against a simple forgetting-based explanation imply that some of the differences in learning between the Blocked and Contrast conditions may have been due to the different contexts in which categories were acquired in the two conditions. Different types of encoding contexts may have caused subjects in the two conditions to form subtly different category representations in memory. In the Blocked condition, categories were learned in relative isolation. In the Contrast condition, instances of previous categories continued to be practiced as new categories were being learned, so each new category was acquired in a context in which category comparisons were required to classify each training instance. It seems reasonable to assume that such frequent comparisons led to representations which differentiated each category from related categories in a more explicit manner than in the Blocked condition. If the representations formed in the Contrast condition highlighted features promoting discrimination more than those formed in the Blocked condition, they should transfer better to a Mixed test block such as that shown at the end of training.

Thus, the fact that category B was learned separately from the other categories in the Blocked sequence might have been partly responsible for the apparent loss of B learning in the test block. Because of this isolated learning context, subjects failed to learn category B in a way that would later allow it to be easily distinguished from other, similar, categories in a mixed testing sequence. The same would have been true of their earlier learning of the A1 and A2 categories. As a result, presenting these categories together at the end of training caused learners to become confused, perhaps mixing up which default values were associated with the different categories, or evoking inappropriate categories to describe particular training instances and thus contaminating the default norms for one category with values from a different category. If subjects' norms for category B were to be contaminated or eroded by such confusion, the sharp drop-off in learning of this category during the test block would be explained.

Transfer of learning

We expected to find patterns of transfer in both conditions of this experiment similar to those observed in the Contrast condition of Experiment 1. Since category A2 shared superordinate defaults with the previous category, A1, whereas category B contrasted with A1 and A2 on both sub- and superordinate defaults, we expected to observe different patterns of transfer when A2 vs. B were first encountered and learned.

Consistent with the results of Experiment 1, when the first instance of A2 was presented, STs in the contrast condition increased for the changed subordinate defaults while decreasing slightly (but significantly) for the unchanged superordinate defaults. By contrast, STs increased nonsignificantly for both types of attributes on the first instance of category B compared to the previous instance of category A2.

Similar patterns of transfer were observed during the training phase of the Blocked condition. Here, subordinate STs increased significantly when the first instance of category A2 was shown, whereas STs for superordinates remained unchanged. When the first instance of category B was shown, STs increased significantly for both types of attributes.

In both conditions, subordinate defaults were studied significantly longer than superordinates on the first A2 trial, but there was no difference between them on the first B trial. The differences between super- and subordinate defaults decreased in an orderly fashion over the first 3 instances of category A2 in both conditions, while STs remained the same for both types of attributes over the corresponding instances of category B.

As in Experiment 1, however, the memory data suggested that the transfer of superordinate defaults from category A1 to A2 was incomplete or imperfect. In both the Contrast and Blocked conditions, memory for both superordinate and subordinate attributes decreased on the first instance of A2 compared to the preceding A1 instance. (Recall that complete transfer would imply that memory on this trial would decrease only for subordinates). A similar pattern was shown on the first instance of category B.

To sum up the transfer results, subjects were apparently able to use prior A1 learning to attend more to new subordinate defaults and less to unchanged superordinates while encoding the first few instances of A2, as predicted by the S+C model of category invention. However, the decrease in memory for superordinates at the start of A2 learning suggests that, as in Experiment 1, the transfer of A1 defaults to A2 may have been limited by subjects' rote memory abilities, i.e., by their difficulty in remembering which defaults had been changed from A1 to A2, and which had remained the same.

Experiment 3

Experiments 1 and 2 showed patterns of transfer supporting the normative S+C model of category invention. However, subjects in those experiments showed incomplete transfer of superordinate defaults, perhaps due to their limited ability to remember which default values had been changed in the new category and which remained the same. If these memory limitations were reduced by the use of stimuli that were more easily remembered -- for example, pictures rather than lists of verbal items -- then in theory subjects should produce complete, 100% transfer of superordinate defaults.

Another problem with the transfer evidence from Experiments 1 and 2 is that the switch from category A1 to A2 always occurred prior to the presentation of category B in those experiments. Thus, when subjects encountered the first instances of category A2, the superordinate defaults which they shared with A1 were the only values of these attributes seen in the training instances up to that point (subjects would have seen all four values during the testing phase of earlier trials, however). By contrast, if category B had been shown prior to A1 and A2, then subjects would have seen two different values on both super- and subordinate attributes before encountering their first instance of A2. In theory, transfer between A1 and A2 should be unaffected by these differences in the sequencing of categories, but the results of the previous two experiments leave this assumption untested.

The present experiment differed from Experiments 1 and 2 in two ways. First, it used pictorial stimuli instead of verbal materials. Since much research (e.g., Paivio, 1969, 1971, 1978) has shown that memory for pictorial materials tends to exceed memory for verbal materials, we expected to obtain more complete transfer between overlapping categories in the present experiment than was observed in Experiments 1 and 2. Also due to this superior pictorial memory, and because pictorial materials are generally more distinguishable than list materials, category learning should be more stable than we observed in prior experiments. Thus, even though the categories in the present experiment were presented in a blocked rather than a contrast sequence, less deterioration in performance should occur during the final mixed test block of the present experiment than occurred in Experiment 2.

The features of pictorial stimuli cannot easily be presented and studied one item at a time like a verbal list, so the attribute-listing procedure used to index unsupervised learning in Clapper and Bower (1991, 1994a) was used in the present experiment, with minor changes in task instructions. Subjects were asked to list only the distinguishing features of each instance; since variables distinguish instances within a category whereas defaults do not, the proportion of defaults vs. variables listed on a given trial provided an index of learning analogous to that provided by STs in Experiments 1 and 2.

The present experiment also differed from Experiments 1 and 2 in presenting four categories instead of three. Two of these categories shared one set of superordinate defaults and differed along their subordinate attributes (categories A1 and A2), while the other two categories shared a different set of superordinate defaults and were distinguished by different sets of subordinate defaults (B1 and B2). The categories in this experiment were presented in the sequence A1 - A2 - B1 - B2.

Method

Subjects

The subjects were 15 undergraduate students of Stanford University, participating in partial fulfillment of an Introductory Psychology course requirement.

Procedure

Subjects were tested individually for a single session lasting 60 min. The training instances were realistic line drawings of fictitious insects, presented in a 48-page, 8 by 11.5 in. booklet. Included with this booklet were printed instructions and an agreement that subjects signed to indicate their informed consent to participate. A single training instance (insect picture) appeared on each page, together with brief instructions for the experimental task.

Subjects were instructed to write on each page those distinctive properties of the presented insect that would be useful for telling it apart from other insects of the same general type. Subjects were told to imagine that they were writing their lists to prepare for a later multiple-choice recognition test in which they would be asked to match each list with its corresponding insect from among several distractor items (i.e., other insects from the same test booklet). It was suggested that subjects should list only those properties that would be necessary to identify an insect on such a test, and to omit all non-distinguishing properties. To increase the force of these instructions, subjects were told to imagine they would be charged $0.25 for each feature they listed, but that they would also be charged $1.00 for each item they answered incorrectly on the multiple-choice test.

Subjects were also instructed to look only at the page of the booklet on which they were currently working, and not to look backward or forward at other pages. They were allowed to complete the experimental task at their own pace. Once they had finished, subjects were given a debriefing page that explained the procedures and goals of the experiment, and were allowed to leave.

Materials and Design

The stimuli were line drawings of fictitious insects, all of which shared a common "base" structure (e.g., head, thorax, abdomen) plus eight dimensions of variation (attributes), such as wing shape, abdominal markings, eye color, etc. Each attribute had 2, 4, or 8 discrete values (e.g., wings of different shapes, eyes of different colors), depending on its role in the experimental design.

Subjects were assigned to two different conditions, referred to as the *Blocked* condition and the *Random Control* condition. In the Blocked condition, the stimuli could be divided into four distinct categories based on the correlational structure of the stimulus set. The categories were designed as follows: Category A1 = 111111xx, category A2 = 111222yy, category B1 = 222333qq, and category B2 = 222444rr, where x,y q, and r denote different pairs of values of variable attributes occurring in each of the four categories. Thus, the variables attribute had a total of 8 different values, but only 2^2 x 4 = 16 out of a possible 8^2 = 64 possible combinations were used, so the total stimulus set consisted of 16 training instances, 4 from each category.

In the Control condition, each attribute had the same number of values as in the Blocked condition, but each attribute varied independently of the others, i.e., no features consistently covaried, so there were no categories in the stimulus set for this group. Thus, 3 of the attributes took on 2 different values, 3 others took on 4 values, and the remaining 2 took on 8 values, for a total of 32,768 possible instances. Of these, 16 were presented in this experiment. These were selected such that none of the attributes were correlated within the chosen subset.

For convenience, the training sequence in the Blocked condition can be divided into a training phase and a test phase. (However, subjects were not informed this distinction, and no break or change in procedure occurred between phases). During the training phase, instances from each category were presented separately in four blocks. Ten instances of category A1 comprised the first block, followed by a block of 10 instances of A2, then 10 instances of B1, and finally 10 instances of category B2. Since there were only 4 possible instances of each category, 6 of these were presented twice in the training condition, and the remaining 2 were shown once. The order of instances within a block was arranged randomly, except that no instance was ever shown twice in succession.

Following the four training blocks, a fifth (test) block consisted of 8 instances, 2 from each category, shown in an intermixed sequence. (No two instances of the same category could occur in succession during this test phase). The instances from each category shown in this block were those that had been shown only once in the prior training block.

The order of stimuli within the Control condition was determined by random assignment, with the constraint that the entire stimulus set be shown before the a given instance was repeated.

Results and Discussion

The data consisted of the proportion listed of each of three types of attributes: the 3 superordinate defaults, the 3 subordinate defaults, and the 2 variable attributes. These data are displayed in Figure 5. A larger number of subjects were assigned to the Blocked condition than to the Control condition, because we wanted to explore the detailed pattern of results from the Blocked condition. The Control condition was included only to provide a baseline and to ensure that learning effects in the Blocked condition would be due to the interfeature correlation rather than to the types of attributes having different numbers of values. Thus, the random assignment procedure was adjusted to assign 3 subjects to the Blocked condition for every one assigned to the Control condition; this caused 11 subjects to be assigned to the Blocked condition and 4 to the Control condition.

Insert Figure 5 about here

The main prediction in this experiment was that subjects in the Blocked condition would show strong, stable learning of all four categories and that the patterns of transfer between categories would provide further evidence for the S+C model of category invention.

As expected, rapid learning of all four categories was clearly evident in the Blocked condition. Learning was indexed by subjects' tendency to list variable attributes and to omit superordinate and subordinate defaults from their lists. The average listing of superordinate defaults over the experiment as a whole was 0.04, that of subordinate defaults was 0.16, and that of variables was 0.94. The differences between the default and variable attributes were highly significant throughout the experiment for all four categories.

In this experiment, variable attributes took on more values (8) than did subordinate defaults (4), which in turn took on more values than superordinate defaults (2). On the basis of this difference, 8-valued variables might be expected to be listed more often than 4-valued subordinate defaults, which would in turn be listed more often than 2-valued superordinate defaults. Within the Control condition, the 2-valued attributes were indeed listed less often (at 0.47) than either the 8-valued attributes (at 0.74) or the 4-valued attributes (at 0.72); the latter two did not differ significantly. Thus, the number-of-values factor contributed to a difference even without correlated attributes and categories. However, the patterns differed in the two conditions; whereas in the Blocked condition the listing of variables was in the higher range and that of super- and subordinate defaults was lower, in the Control condition the listing of 4- and 8-valued features were both in the higher range and that of 2-valued attributes was lower.

Direct comparisons between the Blocked and Control conditions indicated that the learning observed in the Blocked condition was not merely an artifact of defaults and variables having different numbers of values. Using the same summary index of learning described above (subtracting the proportion of defaults listed from that of variables listed on each trial), we compared learning between the two conditions. For the four training blocks, the average difference of 0.83 in the Blocked condition was reliably greater than the 0.15 difference from the Control condition. This difference remained highly significant when examined during the test block or when separated by individual categories. [3] Thus, the data showed strong learning and stable retention of categories in the Blocked condition.

In addition to the strong overall learning observed in the Blocked condition, the patterns of attribute listing at the transition points as different categories were introduced implied significant transfer of learning between them. Listings at these transition points increased only for the defaults that changed with the new categories, but the defaults that remained constant continued to be listed at a low level.

When categories differed only in terms of their subordinate attributes (A1 vs. A2, B1 vs. B2), then listings increased only for these changed attributes, while listing of unchanged, superordinate attributes remained roughly constant. Thus, when the first instance of A2 was shown following learning of category A1, listings of subordinate defaults increased greatly compared to the preceding instance of A1, while listings of superordinates and variables remained unchanged. Listing of subordinate defaults decreased following the first instance of A2, returning to their previous level by the end of the A2 block. The same pattern of significant results occurred at the transition between categories B1 and B2.

When both superordinate and subordinate attributes were changed in a new category (from A2 to B1), listing increased somewhat for both types of attributes. This increase was marginally significant averaged over super- and subordinate defaults ($p <$ .10). The modest increases in listing at this transition may have reflected the fact that six default values were switched rather than only three, as at the other category transitions in this experiment. This added competition may have reduced the number of defaults at each level that might otherwise have been listed.

As expected, listings of both types of attributes declined to previous levels over the next few B1 trials. (Listing of subordinates actually fell to slightly below their level in the previous A2 block, $p < .10$). Listing of variable attributes was unaffected by the switch from category A2 to category B1, remaining near ceiling levels for the attribute listing measure and not differing significantly from listings during the preceding block (p > 25).

Recall that instances of all four categories were shown during the final 8 trials. Listings showed no significant change during this block from those at the end of the previous B2 block ($p > .10$). Thus, the category norms acquired during the earlier blocked training phase were sustained into a mixed presentation test with no apparent reduction in learning.

The results of the present experiment showed no evidence of unstable category learning, as had been observed in the Blocked condition of Experiment 2. The greater retention of category learning observed in this experiment may have been fostered by subjects having better memory for individual instances. This advantage might have been due to the use of pictures rather than as verbal lists, since pictures are typically remembered better than verbal materials (e.g., Paivio, 1969, 1971, 1978). This higher memory for individual instances might have resulted in stronger, more stable norms being acquired for each category. A related point is that pictures are more obviously dissimilar than are equivalent verbal lists; for example, many global or configural features can distinguish different pictures, whereas lists differ primary in the meanings of their individual items. Similarly, common patterns abstracted from several pictorial stimuli might be more distinguishable than otherwise comparable patterns abstracted from a series of verbal lists, over and above the superior memory for individual instances expected for pictorial stimuli. Because of the greater inherent discriminability of pictures and pictorial patterns compared to lists of verbal items, subjects learning pictorial (compared to verbal) categories might be expected to maintain the strength of the category norms even as the context is switched from a blocked to a mixed training sequence.

General Discussion

These experiments investigated unsupervised learning in stimulus domains characterized by partially overlapping categories and subcategories, and aimed to provide information about how categories are discovered and applied to facilitate the learning of further categories and instances within such domains. As in previous research (Clapper and Bower 1994a, 1994b), the experiments provided evidence that correlation-based categories are unlikely to be acquired solely through direct strengthening of correlational rules or associations; rather, our evidence strongly suggested that categories were invented explicitly in response to the contrast between novel stimuli and the norms of already learned categories.

Evidence for such contrast-based learning had previously been obtained for different stimulus modalities (pictorial vs. verbal stimuli), and for two different task paradigms with a total of three dependent measures (listing preferences in the attribute listing task, STs and recognition memory in the instance memorization task). Contrast-based learning was also implicated when default values occurred probabilistically as well as in 100% of the instances from a given category.

The present results provide additional evidence for the generality of a category invention process: we show that it is not restricted to two category situations investigated before, but also characterizes unsupervised learning when a larger number of contrasting patterns are present in the stimulus set. Moreover, the present results demonstrated subjects' use of category invention to acquire new categories that shared default values with prior categories (i.e., in a hierarchy) whereas in our previous research the categories had contrasting values on all their default attributes.

The combined evidence from this research suggests that category invention is a central method by which human learners acquire categories based on correlational patterns observed in unsupervised tasks. While these experiments do not deny the possibility of autocorrelation as a learning method in some situations, our results do constitute a strong "existence proof"for category invention as a basic capability of human learners for internalizing such patterns.

The present experiments also demonstrated significant transfer of default norms between overlapping categories, and provided evidence consistent with the normative S+C model of category invention. Thus, people used existing category knowledge to increase their efficiency of learning contrasting categories, much as category knowledge is used to improve the learning of individual training instances. The evidence for the transfer of shared defaults was quite robust, occurring for both verbal and pictorial stimulus materials, with different numbers of categories (3 vs. 4), and at different points in the training sequence (early vs. late).

Despite this evidence for the transfer of shared (superordinate) defaults, the memory data from Experiments 1 and 2 implied that this transfer was not complete or flawless. Subjects' recognition memory for shared defaults decreased for the first instance of a new category, whereas in theory performance should have remained at the same level as in the preceding reference category. One plausible explanation of such incomplete transfer is simple memory limitations, i.e., subjects during memory testing may have been unable to remember the full list of changed default values after seeing the first instance of the new category, leading to uncertainty about which 7 values of the reference category had been changed and which 3 had remained the same for the new category.

Importantly, this uncertainty would have only been present during the testing phase of a given trial, but not during the initial study phase. During the study phase, subjects should have had no difficulty in distinguishing between old and new defaults, since the new "defaults" would at that point appear as salient violations of the subjects' existing category norms. But once the training instance had disappeared from subjects' computer screens and the memory testing of that instance began, subjects' lack of complete memory for the features shown during the study phase could then reduce their confidence on the memory tests.

The experiments described above also suggested that certain factors may limit or constrain subjects' retention of earlier categories in the face of decay and interference caused by learning later categories. In Experiment 2, subjects who learned categories separately in a blocked sequence showed significant loss of this learning when the categories were later shown in a mixed sequence. The reduction in later performance may have arisen because subjects became confused as to which particular default values were associated with which category, mixing up default values from different categories and eroding their confidence in the defaults of all the categories.

Such instability was less evident in the Contrast conditions of Experiments 1 and 2. These subjects learned each category in the context of prior categories, i.e., instances of previously-learned categories continued to be shown intermittently as the new category was learned. Thus, subjects received practice in distinguishing each new category from similar, previously learned categories. Such training presumably caused them to form category representations that highlighted the useful distinctions between the different categories, and which thus facilitated effective discrimination performance. On the other hand, subjects who learned each category separately in the Blocked condition of Experiment 2 might have formed category representations that were more fragile than those formed in the Contrast condition, and these differences caused poorer discrimination performance in the later mixed sequence. The advantage of our Contrast condition is much like the long-term memory advantage conferred by distributed rather than massed practice of verbal associations (see, e.g., Baddeley, 1990).

Experiment 3 employed pictorial stimuli which led to very stable categories in the Blocked condition of that experiment, compared to the verbal categories of the Blocked condition of Experiment 2. Stimulus materials that are inherently easier to remember and distinguish should lead to categories (or category norms) that are more memorable and distinguishable. This generalization should hold true not only for the difference between pictorial and verbal stimuli, but also for other stimulus factors affecting memorability such as complexity, familiarity, and meaningfulness of the training stimuli.

Theoretical implications

A variety of models have been published in the literature of cognitive psychology and artificial intelligence that address the problem of learning categories without supervised feedback. The present results, taken together with those of Clapper and Bower (1991, 1994a, 1994b), place strong constraints on which of these can be considered accurate descriptions of unsupervised learning by humans when memory is greatly limited. In particular, an accurate model must include a capacity for inventing new categories based on explicit contrast with the norms of prior categories. Models restricted to some form of autocorrelation (such as testing a series of correlational hypotheses or accumulating a matrix of interfeature correlations) would seem unable to accommodate the pronounced contrast effects observed in these experiments. Of course, such contrast effects do not imply that humans lack any capability for autocorrelation, nor do they rule out a mixed model which includes combinations of category invention and autocorrelation. However, the low level of learning observed in mixed training sequences of Experiment 1 and in our earlier experiments (Clapper and Bower, 1994) suggests that autocorrelation is a rather weak learning method in humans, at least for the types of categories and complex stimulus materials used here.

The present results also indicate that an acceptable model of human unsupervised learning should have the capability to transfer generalizations learned about previous categories to overlapping new categories. Thus, the category invention process must conform approximately to the basic S+C encoding framework. Models that, in effect, discard all prior norms when an executive process decides to assign a novel stimulus to a new category are thus incompatible with our results. For example, certain 2-layer connectionist models (e.g., Grossberg, 1980; Carpenter & Grossberg, 1987; Rumelhart & Zipser, 1986) include a capacity for "novelty detection": If a pattern of activity on the first ("stimulus") level of the network is sufficiently different from previous patterns, a new second-level ("category") unit is activated. This is equivalent to assigning the novel stimulus pattern to a new category. However, associations between shared pattern elements and the first response unit (category) are not transferred to the second response unit in such a network, thus such a model would not produce the default transfer observed in the present experiments.

We also think that an adequate learning model should include the capability to learn categories while at the same time using that knowledge to optimize the encoding of individual training instances. This optimal encoding is achieved by allocating more attention to the features of an instance the more uncertain or novel the feature is with respect to category norms. This pattern of attentional allocation was observed in all our instance memorization experiments, as well as in earlier experiments investigating schema-based memory for events and persons (e.g., Graesser et al, 1980; Srull & Wyer, 1989). Most models designed to learn categories or correlational patterns assume that subjects are biased to attend mainly to features that are diagnostic (predictive) of the categories being learned; these predictive features are the same as our category defaults (e.g., Billman & Heit, 1988). Such models are designed mainly to classify instances, not to apply them to facilitate further learning (i.e., of instances or of related categories), and so they postulate patterns of attentional allocation that are contrary to those expected by the S+C encoding model. The attentional assumption of such models would need to be revised to accommodate the patterns of attentional allocation observed in our unsupervised learning experiments.

In addition to providing evidence for basic capabilities such as category invention or default transfer, the present experimental paradigms provide rich information about additional factors which control or limit the application of these basic capabilities. For example, we have seen how the stability of categories can be limited by factors relating to the stimulus materials or the context of acquisition. By accumulating information about how various factors influence unsupervised learning, such research should enable the development of more powerful and detailed theoretical models, with accompanying increases in predictive power. By observing when people seem to conform to normative learning strategies, and when human capacity limitations or other factors interfere with such normative strategies, we may begin to construct a detailed picture of human learning mechanisms.

References

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge, & S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension.* Hillsdale, NJ: Erlbaum.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413-451.

Baddeley, A. (1990). *Human memory: Theory and Practice.* Boston: Allyn & Bacon.

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, *12*, 587-625.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*, 177-220.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54-115.

Clapper, J. P., & Bower, G. H. (1991). Learning and apply category knowledge in unsupervised domains. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory. Vol. 27.* New York: Academic Press.

Clapper, J. P., & Bower, G. H. (1994b). Instance and category learning in unsupervised tasks. *Submitted for publication*, .

Collins, A., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Collins, A., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240-247.

Davis, B. R. (1985). An associative hierarchical self-organizing system. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*, 570-579.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 234-257.

Graesser, A. C., Woll, S. B., Kowalski, D. J., & Smith, D. A. (1980). Memory for typical and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 503-513.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1-51.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery.* Cambridge, MA: MIT Press.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory. Vol. 18.* New York: Academic Press.

Kolodner, J. L. (1984). *Retrieval and organizational strategies in conceptual memory: A computer model.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach.* Palo Alto, CA: Tioga Publishing Company.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482-488.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychology Review*, *76*, 241-263.

Paivio, A. (1971). *Imagery and verbal processes.* New York: Holt, Rinehart & Winston.

Paivio, A. (1978). Dual coding: theoretical issues and empirical evidence. In J. M. Scandura, & C. J. Brainerd (Eds.), *Structual/process models of complex human behavior.* Nordhoff: Leiden.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1* (pp. 45-77). Cambridge, Mass.: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery in competitive learning. In J. L. McClleland, & D. E. Rumelhart (Eds.), *Parallel distributed process: Explorations in the microstructure of cognition, Vol. 2*. Cambridge, Mass.: The MIT Press.

Schank, R. C. (1982). *Dynamic memory.* Cambridge, UK: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Hillsdale, N. J.: Lawrence Erlbaum Associates.

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review*, *96*, 58-83.

Footnotes

This research was supported by a grant (AFOSR-91-0144) from the Air Force Office of Scientific Research and a grant (MH-47575) from the National Institute of Mental Health.

Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, CA 94305.

[1] In proposing this dichotomy, we exclude models which assume that learners first memorize an entire set of training instances, and then compute an optimal set of generalizations across this memory set (e.g., Fried & Holyoak, 1984; Michalski & Stepp, 1983); Such models seem unrealistic as descriptions of human learning due to their assumption of unlimited memory and computational capacity. In this article, we assume that subjects update their conceptual norms in response to each instance they encounter (sequential or incremental learning assumption). This means that on each trial subjects compare the current instance to their existing category norms, and update these norms based on this simple comparison.

[2] We may conceive of subjects inferring default norms by a Bayesian strategy in which they begin with prior beliefs about equiprobability of values of an attribute, and they modify these equiprobable-value beliefs as successive instances exhibit consistent values (of 1 or 2) on the relevant attributes. Thus, subjective confidence in default norms would reflect this pooling of instance data with prior beliefs.

[3] Since the Control condition presented a random sequence throughout the experiment with no division into different blocks, we used the difference averaged over all 48 trials as our reference baseline for these comparisons.

Figure Captions

Figure 1. Sample stimulus sets illustrating how categories are defined in terms of correlated attribute values. The 8 attributes are arrayed in columns, and numerals (1, 2) denote different values of the column attribute. An x denotes a variable attribute which may take on either value for instances with that category. Fig. 1a illustrates two categories defined by 5 perfectly correlated features. Fig. 1b illustrates patterns with no discernible categories.

Figure 2. Computer display as it appeared during each phase of Experiments 1 and 2.

Figure 3. Study time and recognition accuracy data from Experiment 1. In this figure, the functions connecting the "O" points are for the superordinate defaults, those connecting the "*" points are for the subordinate defaults, and the "." points are for the variable attributes. The plots are divided by category, and the plots for the Contrast condition are further subdivided in terms of whether a given set of instances occurred in the first, second, or third block of trials.

Figure 4. Study time and recognition accuracy data from Experiment 2. As in Figure 3, the "O" points are for superordinate defaults, the "*" points are for subordinate defaults, and the "." points are for variable attributes.. For the Contrast condition, the plots are divided by category and block as in Figure 3. For the Blocked condition, the plots are also divided by category, and the trials for each category are separated into training vs. test blocks.

Figure 5. Attribute listing data from Experiment 3. Trials are shown in their original order in both plots, and the plot for the Blocked condition is divided by categories and to indicate the mixed test block.

Figure 1.

a)

Attribute							
1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	2
1	1	1	1	1	1	2	1
1	1	1	1	1	1	2	2
1	1	1	1	1	2	1	1
1	1	1	1	1	2	1	2
1	1	1	1	1	2	2	1
1	1	1	1	1	2	2	2

Attribute							
1	2	3	4	5	6	7	8
2	2	2	2	2	1	1	1
2	2	2	2	2	1	1	2
2	2	2	2	2	1	2	1
2	2	2	2	2	1	2	2
2	2	2	2	2	2	1	1
2	2	2	2	2	2	1	2
2	2	2	2	2	2	2	1
2	2	2	2	2	2	2	2

Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

b)

Attribute							
1	2	3	4	5	6	7	8
1	2	2	2	2	2	1	1
1	2	2	1	2	1	2	2
2	1	2	2	2	2	1	2
1	1	2	2	1	1	2	1
2	2	2	2	1	1	1	2
2	1	1	2	1	2	2	1

Attribute							
1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	2
2	2	2	1	2	2	2	2
2	1	1	1	2	2	2	1
2	2	1	1	2	2	2	1
1	1	1	1	2	1	1	2
1	2	1	2	1	2	1	2

No categories defined

Figure 2.

```
a.      Aralia

        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        dark grey bark
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
        XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

        Press INS or DEL to see next item
```

```
-------------------------------------------------------------
```

```
b.      Aralia

        1. deep brown bark
        2. dark grey bark
        3. mossy green bark
        4. light tan bark

        ******************************
        * Enter a number from 1 to 4 *
        ******************************
```

```
-------------------------------------------------------------
```

```
c.      Aralia

        1. deep brown bark
    --> 2. dark grey bark
        3. mossy green bark
        4. light tan bark

           INCORRECT

    Arrow indicates correct choice

        Press RETURN to go on
```

Figure 3.

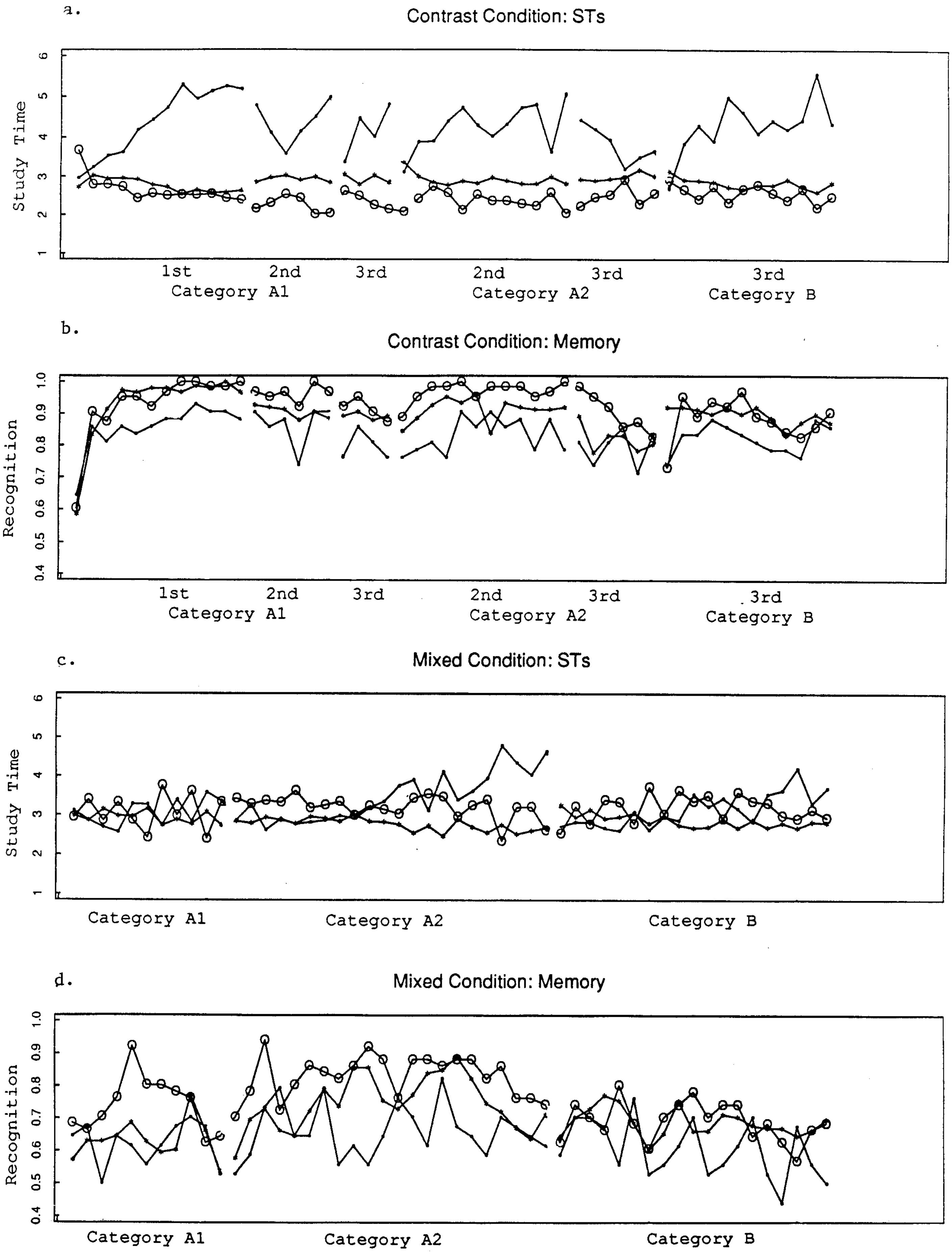
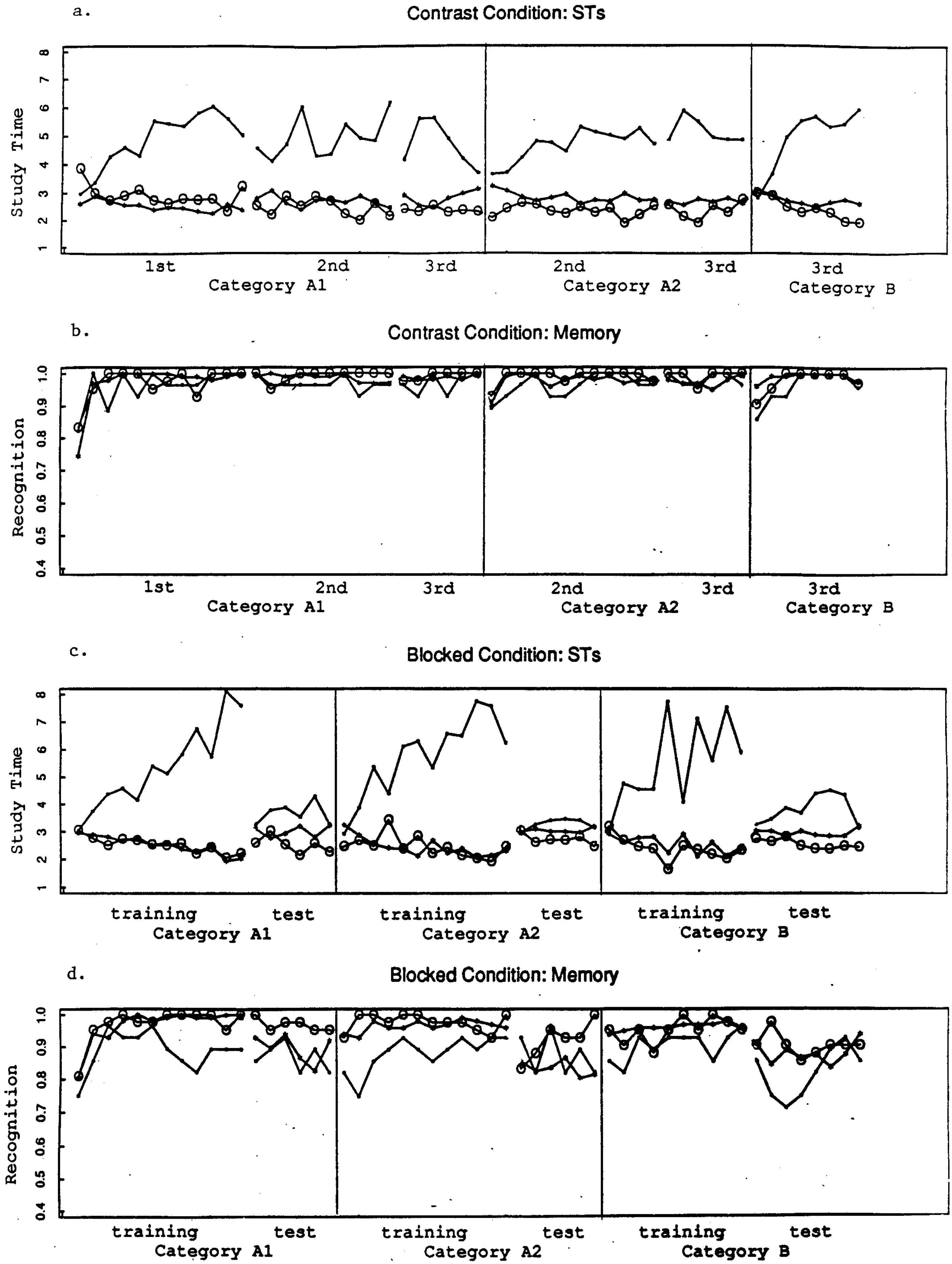

a.
Contrast Condition: STs
Study Time
1
2
3
4
5
6
1st
2nd
3rd
2nd
3rd
3rd
Category A1
Category A2
Category B
b.
Contrast Condition: Memory
Recognition
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1st
2nd
3rd
2nd
3rd
3rd
Category A1
Category A2
Category B
c.
Mixed Condition: STs
Study Time
1
2
3
4
5
6
Category A1
Category A2
Category B
d.
Mixed Condition: Memory
Recognition
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Category A1
Category A2
Category B

Figure 4.



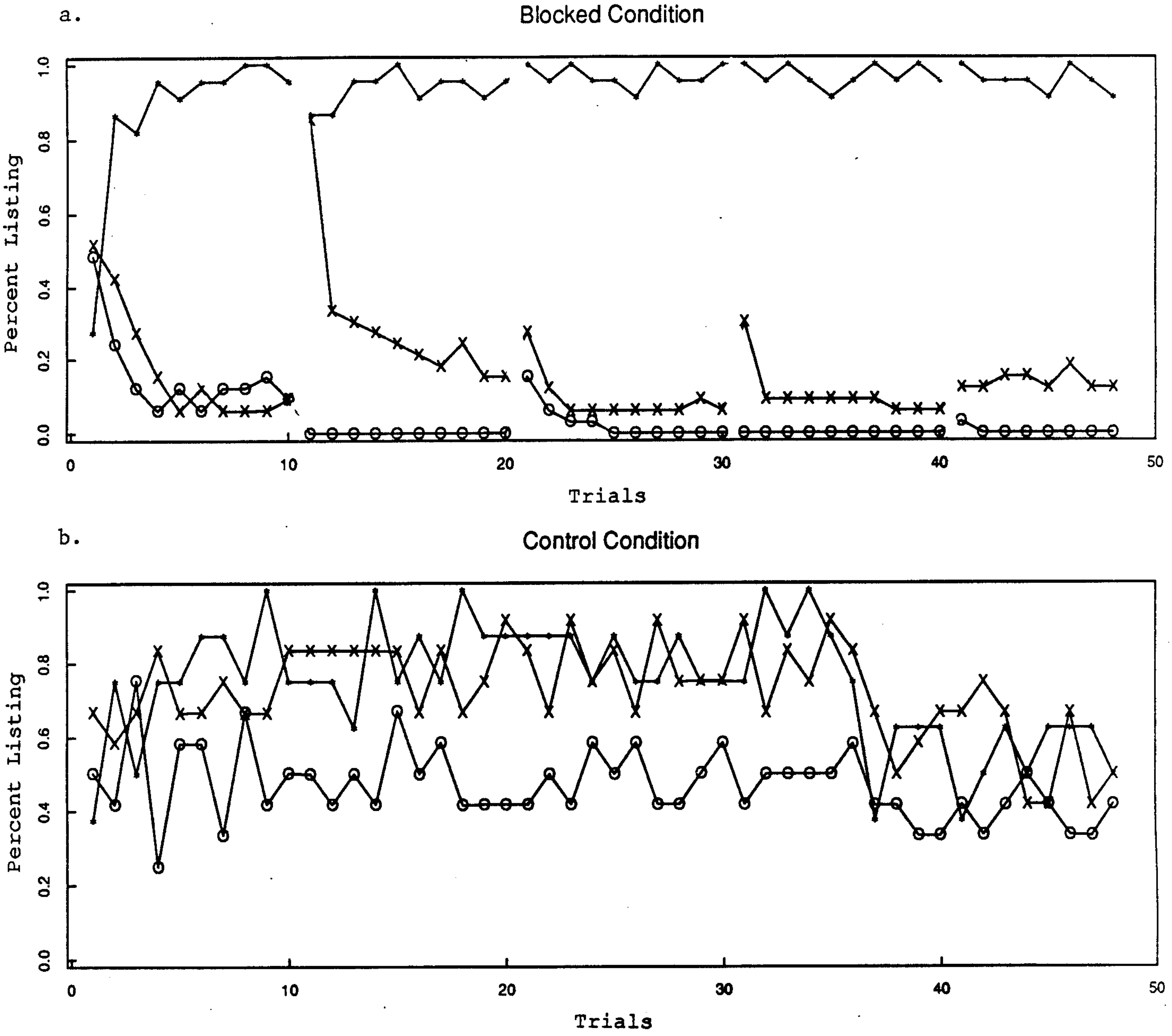
a.

b.

c.

d.

Figure 5.


a.
Blocked Condition
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Trials
b.
Control Condition
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Trials

Learning Categories Without Teachers

By

Gordon H. Bower and John Clapper
Stanford University

Paper presented at "Practical Aspects of Memory" Conference. University of Maryland. August 2, 1994
Research Supported by the Air Force Office of Scientific Research, Grant #AFOSR-87-0282

One of the practical tasks people face is how to learn about their environment, in particular, how to categorize and classify the objects and events around them. Practically all experimental research on category learning has studied what is called "supervised learning," wherein a tutor or supervisor teaches a concept to a learner by providing trial by trial feedback regarding the learner's tentative classification of a series of patterns.

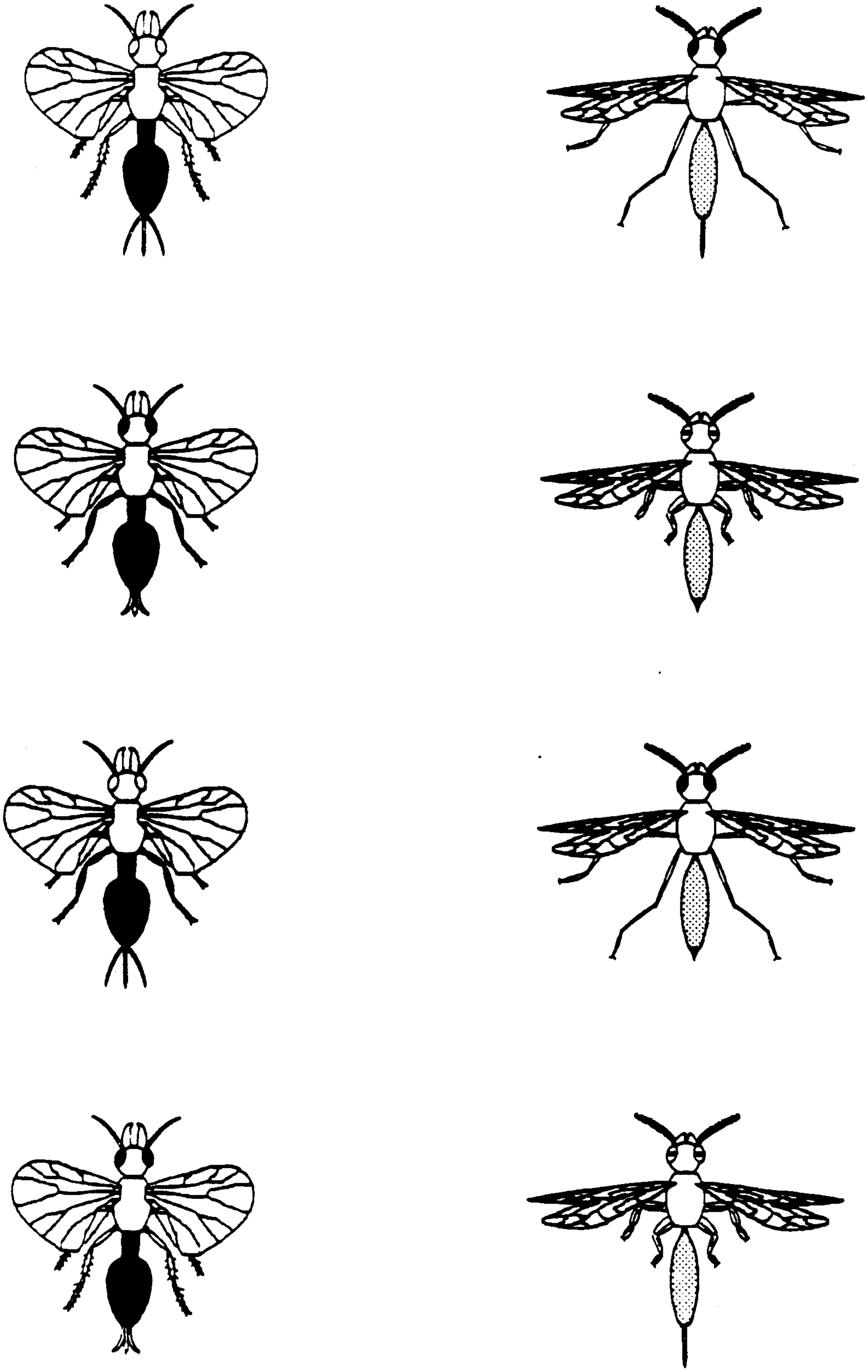
We will address a different issue here, namely, how people learn categories when they have no teacher, when left on their own to discover any usable clustering of stimuli that they can. We call the general paradigm "unsupervised learning", because it involves no supervisor or trainer who provides feedback to learners about the current classification. In fact, in our experiments with college students, we never mention categories or category learning. From the subjects' point of view, they are simply trying to memorize each instance or stimulus pattern as it's presented.

We believe that this kind of unsupervised discovery of categories occurs often in real life, perhaps as preverbal children explore their world of perceptual objects, as they learn their language, or whenever pioneers in any unexplored field try to classify the varieties of things that nature serves up to them. Unsupervised category learning occurs in formal school settings often under the name of "discovery learning" wherein students are permitted to explore a given physical domain in hopes that they will stumble upon its underlying structure or principles.

In our research we use as stimuli collections of trees or of insects such as these (Overhead #1). We've composed these pictures on a Macintosh computer; the insects vary in many physical features, and we define a category of bugs according to which features go together. Thus, these bugs arranged in two columns fall into two categories; they differ in their body shape, color, type of wings, antennae, and front pincers, whereas the eyes, tails, and legs vary within the categories. We can represent the stimuli in abstract binary notation with each instance represented as a row vector, as shown here (Overhead #2). In this illustration, the first 5 of 8 attributes have correlated values of 1 in Category A and 2 in Category B, whereas the last 3 attributes vary randomly within the categories. We refer to the first attributes as predictable or default values of the relevant attributes, and the second as variable values of the unpredictable attributes.

We were interested in two questions. The first question was how to measure category learning in this domain where categories are never mentioned to subjects. The second question was how to arrange sequences of training stimuli shown one at a time in order to speed up category discovery and use.

= 1



#2

Attribute							
1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	2
1	1	1	1	1	1	2	1
1	1	1	1	1	1	2	2
1	1	1	1	1	2	1	1
1	1	1	1	1	2	1	2
1	1	1	1	1	2	2	1
1	1	1	1	1	2	2	2

Attribute							
1	2	3	4	5	6	7	8
2	2	2	2	2	1	1	1
2	2	2	2	2	1	1	2
2	2	2	2	2	1	2	1
2	2	2	2	2	1	2	2
2	2	2	2	2	2	1	1
2	2	2	2	2	2	1	2
2	2	2	2	2	2	2	1
2	2	2	2	2	2	2	2

Category "A" : 1 1 1 1 1 x x x
Category "B" : 2 2 2 2 2 x x x

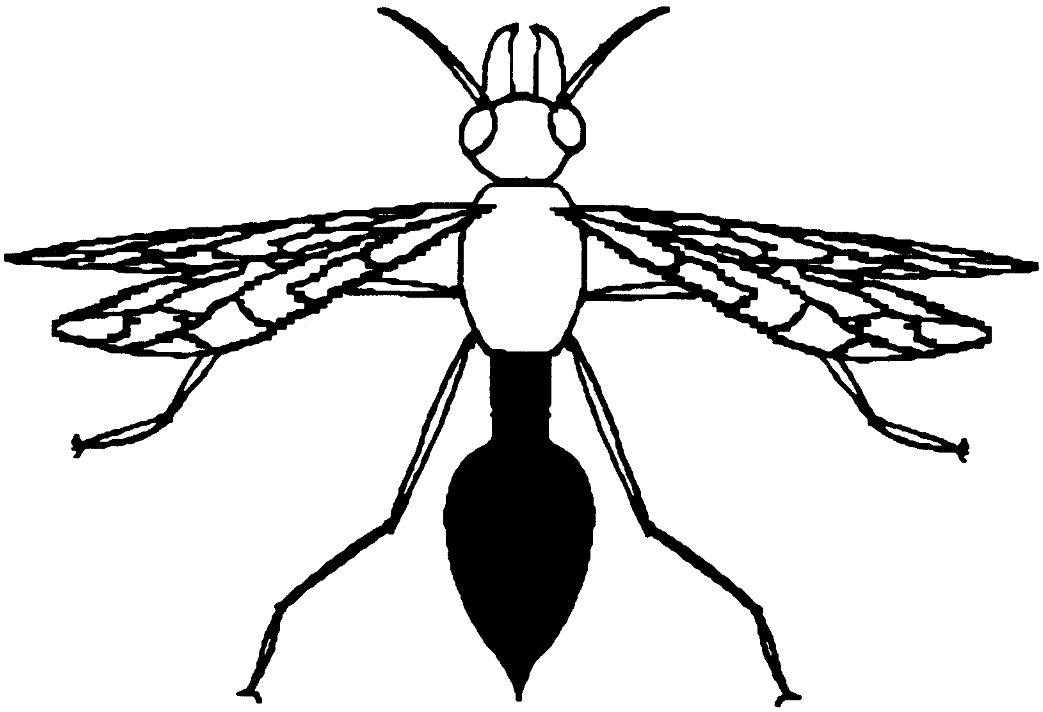
Turning to the first issue, how might we measure category learning in such situations in which categories are never mentioned? We did this by giving subjects the goal of remembering each of the instances as they were presented one by one. Shown a bug, subjects were asked to list a few of its attributes that would enable them to remember it later, in the sense of being able to pick it out from 3 other similar bugs on a recognition memory test. Subjects were urged to list as few attributes as possible, but ones that would be maximally informative for picking out this bug on a later test. That later test in fact was never given.

Here is an example (Overhead #3) of what a subject listed for this insect: white eyes, thin wings, black vase-shaped abdomen, 1 stinger, 2 white antennae, and 2 clawlike pincers.

So, what might we expect subjects to do in this task? Subjects might play dumb, and simply list and record in memory for each instance most of its features, leading to a memory representation like that shown at the top of this overhead (Overhead #4). But an intelligent learner ought to follow a "schema-plus-corrections" strategy which leads to a type of memory representation illustrated at the bottom of the overhead. By this strategy, subjects should record a new instance by first noting one or more of its defaults to indicate its category, and then, by listing the unpredictable or variable attributes which would serve to uniquely identify this particular instance within the category. Of course, to follow this strategy, subjects must have first formed a category based on noticing consistently correlated defaults. Looked at from another perspective, however, we can take as an indirect measure of category learning the extent to which subjects stop listing the predictable defaults but increasingly list the unpredictable, variable attributes of the instances.

It turns out that actual learners--at least, college students--come to approximate this ideal pattern. Their performance is illustrated in the next overhead (Overhead #5); these are subjects who first see 16 patterns of one category (call it A) followed by 16 of a second category (called B), followed finally by a mixture of 4 A and 4 B test patterns. The top figure shows the percent of default attributes that subjects list; this drops rapidly from around 60% for the first bug to around 15% by the 10th bug. The listing of default attributes rises abruptly when the first B pattern comes along, since subjects are surprised by the novel values of its default attributes. But here again, listing of default values for the B patterns quickly drops off to near the minimal value as the B-category norms are learned. The minimum number of defaults subjects should record is one out of 5, or 20%, which is exactly where they're performing. It's significant, too, that subjects are not disrupted when, in the final block of trials, they encounter a mixed series of A's and B's. They obviously were able to maintain intact the separate norms for the two categories.

#3



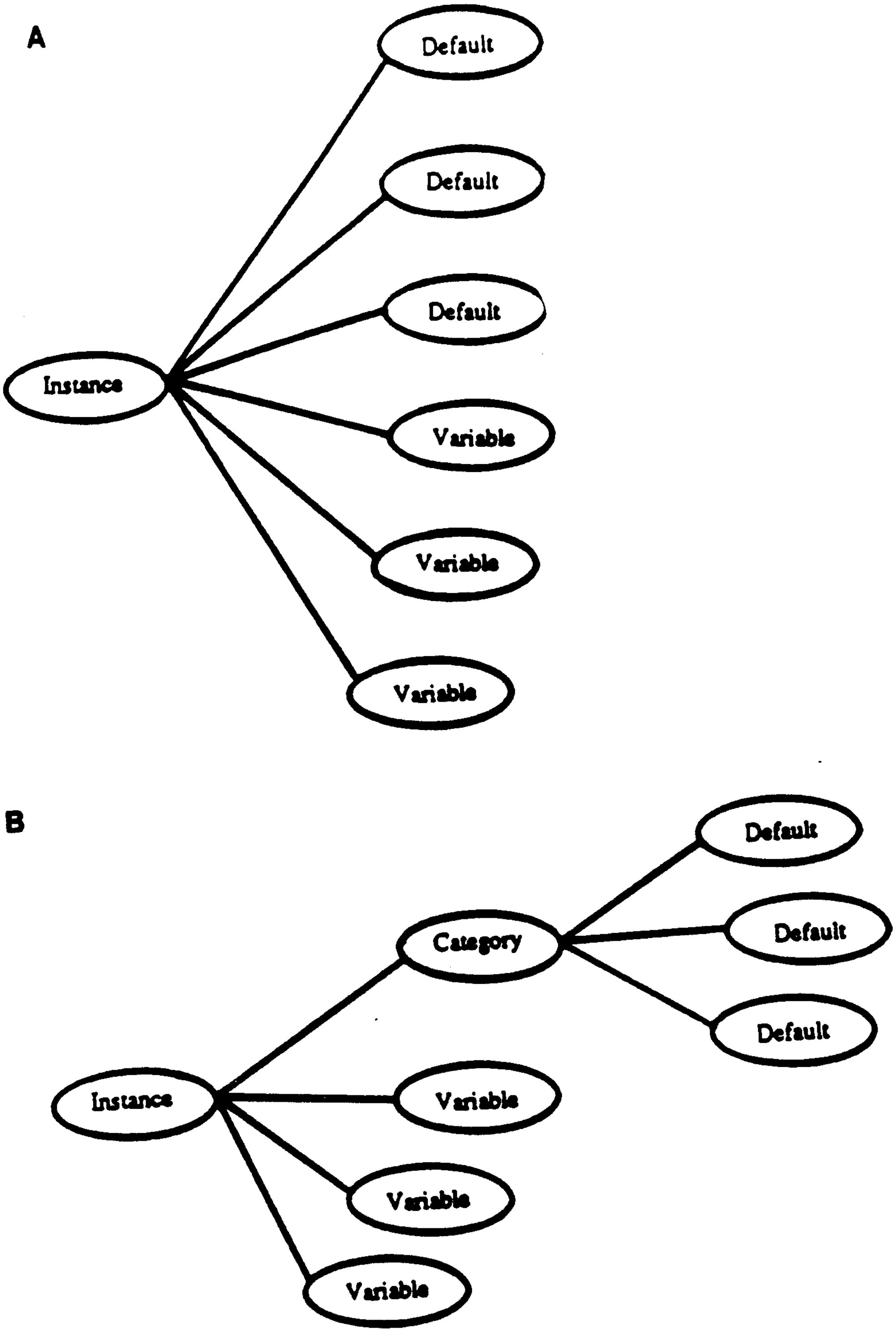
1
1
0
1
1
1
1
1

DO NOT WRITE OR MAKE ANY MARKS ABOVE THIS LINE

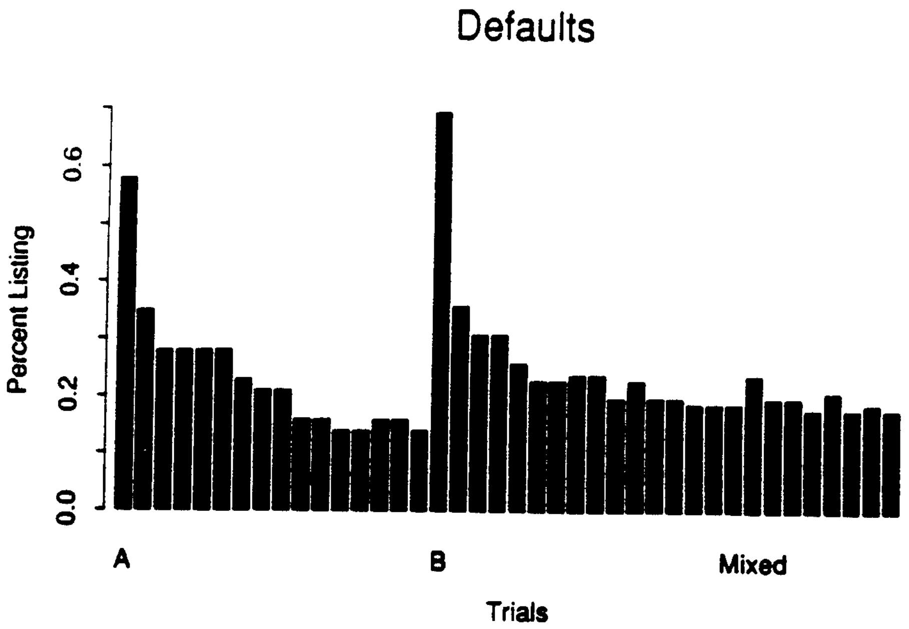
Write down a short list of this insect's features. List ONLY those features that you'd need to identify this insect on a later multiple-choice test. Imagine that: (1) each feature on your list will cost you **25 cents**, and (2) each misidentification on the multiple-choice test will cost you **1 dollar**.

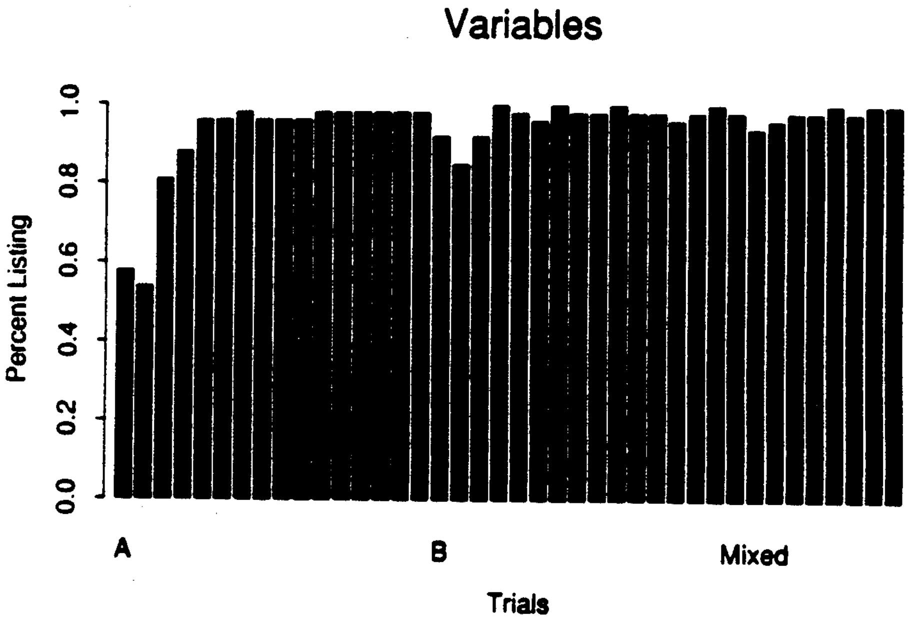
1) White eyes
2) 4 thin wings
3) Black vase shaped abdomen ending in
   1 stinger
4) 2 longwhite antennae + 2 long white clawlike ones

DO NOT LOOK AT ANY OTHER PAGES.


A
Default
Default
Default
Instance
Variable
Variable
Variable
B
Default
Category
Default
Default
Instance
Variable
Variable
Variable

#5


Defaults
Percent Listing
0.0
0.2
0.4
0.6
A
B
Mixed
Trials

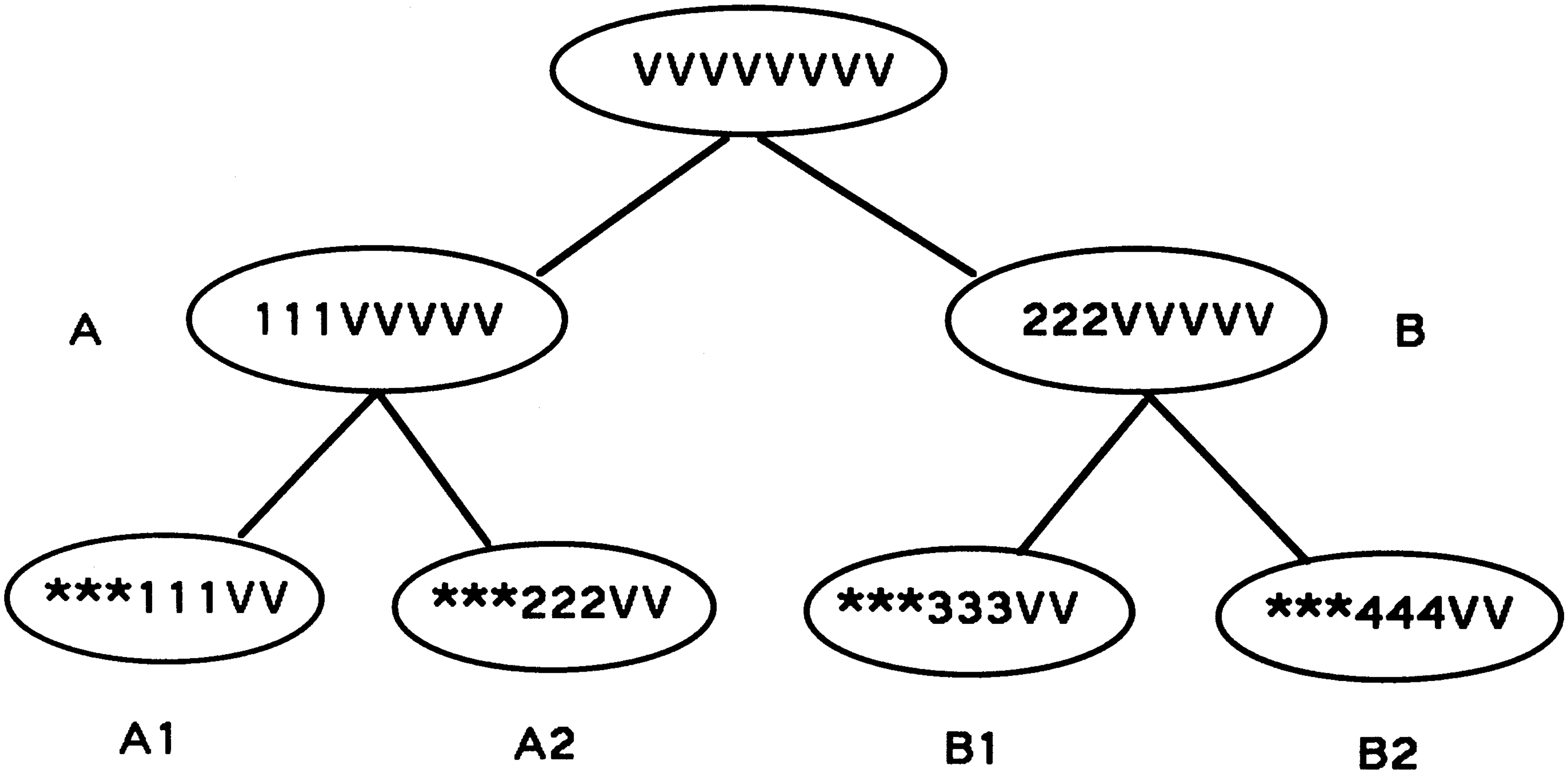

Variables
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
A
B
Mixed
Trials

The percent listing of unpredictable, variable attributes at the bottom of the overhead is just the mirror image reflection of the default-listings above. During the inital A-series, subjects learn to identify and list increasingly the variable attributes of the instances; this dips a bit when they hit the first surprising B stimulus when they have to record the new defaults, but listing of variables quickly recovers to a high level throughout the following mixed series of As and Bs. As you can imagine, a useful index of learning is simply the difference in percentage listing of variable minus default attributes.

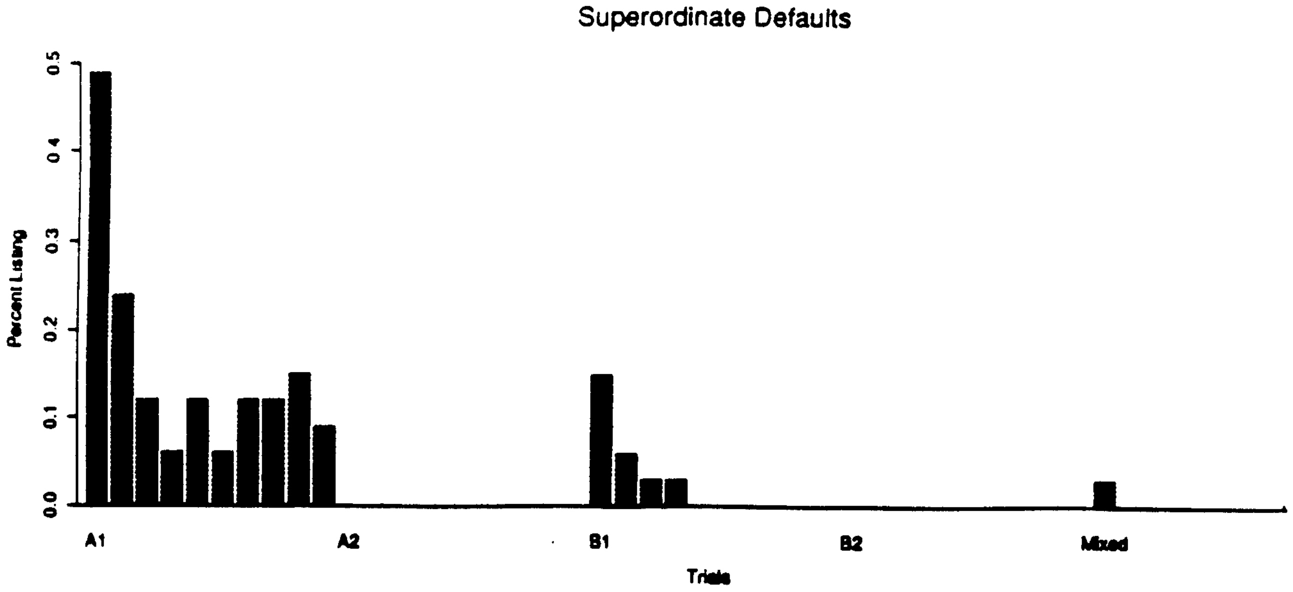
This same procedure was used to follow subjects learning a hierarchy of nested categories, which is illustrated here (Overhead #6), where V stands for a variable attribute. The first two categories, A1 and A2, share values of 1 on the first 3 attributes but differ with values of 1 versus 2 on the fourth, fifth, and sixth. The last two categories, B1 and B2, share values of 2 on the first 3 attributes but likewise differ with value of 3 versus 4 on the fourth, fifth, and sixth attributes. These were shown to subjects in blocks of 10 instances with the blocks in the order A1, A2, B1, B2, all ending with a mixed test block of 2 instances from all 4 subcategories. Here are the attribute listing data (Overhead #7). The top graph shows the rapid decline in listing of the superordinate (the first three) defaults, but it pops up again at the first B stimulus when these values are changed. The middle graph is for the subordinate defaults--attributes 4, 5, and 6; these decline over the first series of A1 patterns, then pop up when the first instance of the novel A2 subcategory is seen, and pop up again as each new subcategory makes its appearance. However, all the superordinate defaults are carried over without fail as new subordinate categories are created to decribe the new instances. By comparison, the fully variable features, shown in the bottom graph, rise up quickly and continue to be listed around 95% throughout. So these data show excellent learning of a hierarchy of categories of insects defined by common features.

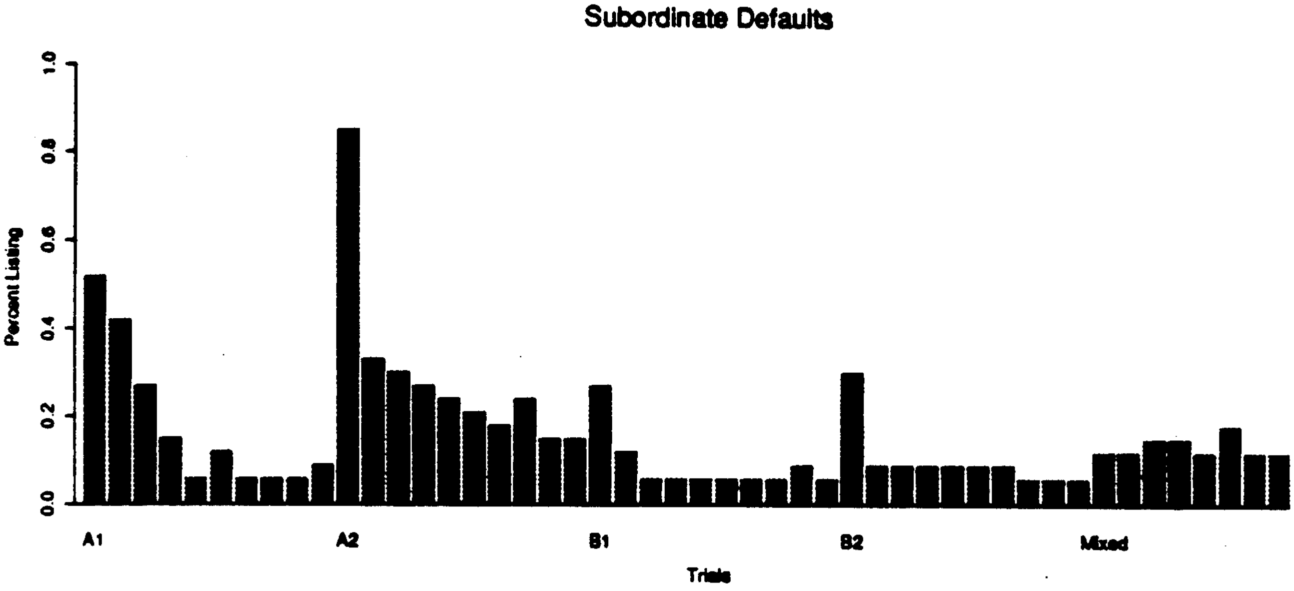
Another indirect index of such learning we have explored involves limiting the amount of time subjects can inspect a verbal description of a species of trees while trying to memorize it (Overhead #8). They are allowed to examine the attributes one at a time on the computer screen. They can move around among the attributes row by row, and we record how much of the time they invest in studying default versus variable attributes. After studying each instance, their memory for it is immediately tested. As expected, subjects quickly learn the defaults, so they spend progressively less time inspecting them, but nonetheless recall them very well; on the other hand, they spend a progressively higher percentage of their time studying the variable attributes, so that their memory for those features also improves. I'll not have time here to show you any of that data on self-selected study times, although it is very orderly and regular.[Overhead off]

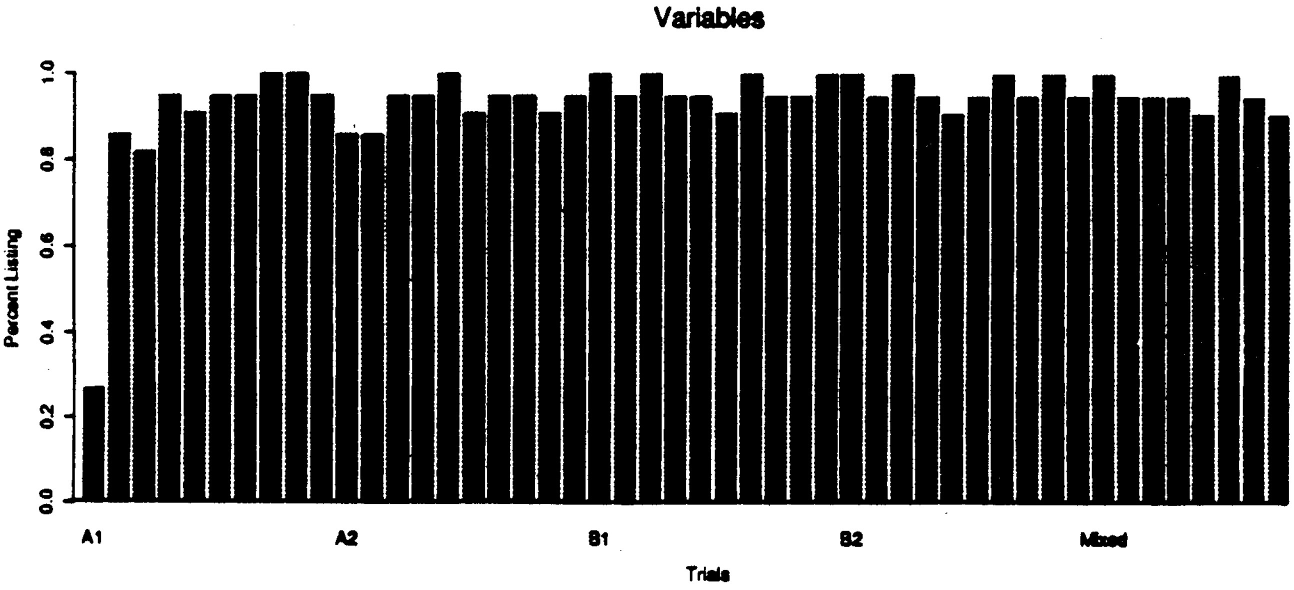
7#


VVVVVVVV
A
111VVVVV
222VVVVV
B
***111VV
***222VV
***333VV
***444VV
A1
A2
B1
B2

#1


Superordinate Defaults
Percent Listing
0.0
0.1
0.2
0.3
0.4
0.5
A1
A2
B1
B2
Mixed
Trials


Subordinate Defaults
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
A1
A2
B1
B2
Mixed
Trials


Variables
Percent Listing
0.0
0.2
0.4
0.6
0.8
1.0
A1
A2
B1
B2
Mixed
Trials

II8

a.

```
Aralia

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
dark grey bark
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Press INS or DEL to see next item
```

---

b.

```
Aralia

1. deep brown bark
2. dark grey bark
3. mossy green bark
4. light tan bark


********************************
* Enter a number from 1 to 4 *
********************************
```

---

c.

```
    Aralia

    1. deep brown bark
--> 2. dark grey bark
    3. mossy green bark
    4. light tan bark


          INCORRECT

Arrow indicates correct choice

      Press RETURN to go on
```

I will turn now to describing a procedural variable we've studied which has a major impact on how much subjects learn from exposure to a set of A and B patterns. This variable is simply the sequence or order in which the two sets of patterns are presented to subjects. We were not prepared for the huge effect this variable had on our data, namely, that category learning is far, far easier if subjects see a long block of instances all of one type before they ever see instances of a second type. Compared to a randomly intermixed series, the advantage in learning produced by blocking is just enormous.

You can get an intuitive feel for the difference in difficulty here by perusing the first 10 instances (in the rows) which mix 5 A-patterns with 5 B-patterns in random order in this overhead (Overhead #9); your job is to find out what values of which attributes are correlated. It's very hard to do; moreover, our subjects weren't allowed to examine all 10 patterns laid out before them at one time as you can see here, they could only see one instance at a time. Iin addition, our subjects did not have the goal of looking for correlated features underlying categories.

The structure of the patterns becomes more obvious and easily learnable if the same 10 instances are presented in a block of 5A's, then 5B's, as illustrated at the bottom. Here, looking down the columns, you can see that the A's have a value of 1 in attributes 2, 5, 7, and the B's have a value of 2 in those attributes.

It's easy to demonstrate this principle, that category learning is facilitated by blocking instances. One demonstration is shown in this overhead (Overhead #10) comparing learning of two groups after they were pretrained with 8 instances. In pretraining, the Practice group (the Xs) saw 4 A's mixed in with 4 Bs, whereas the Contrast group (the O's) saw a block of 8 A's. These two groups then received a mixed series of 12 A's and 12 B's, but we've plotted the A and B trials separated in this graph. The learning measure is the subjects' preference for listing variable over default attributes.

It's obvious that the Contrast subjects, who start out seeing enough A's in a row to acquire very confident norms, continue to do well even after they are surprised by the first B pattern; they consistently outperform the Practice subjects, whose pretraining with a mixture of A's and B's seems to have locked them into a very nonoptimal performance, barely above that of a random uncorrelated control condition. Paradoxically, the Contrast condition shows that by eliminating the B trials in Pretaining, we enhance the learning of B stimuli in the later series.--- sort of a reverse practice effect. The Practice subjects apparently see so much variability in their pretraining mixture of A's and B's that they give up any attempt to see structure in the collection of patterns. [Overhead off]

4

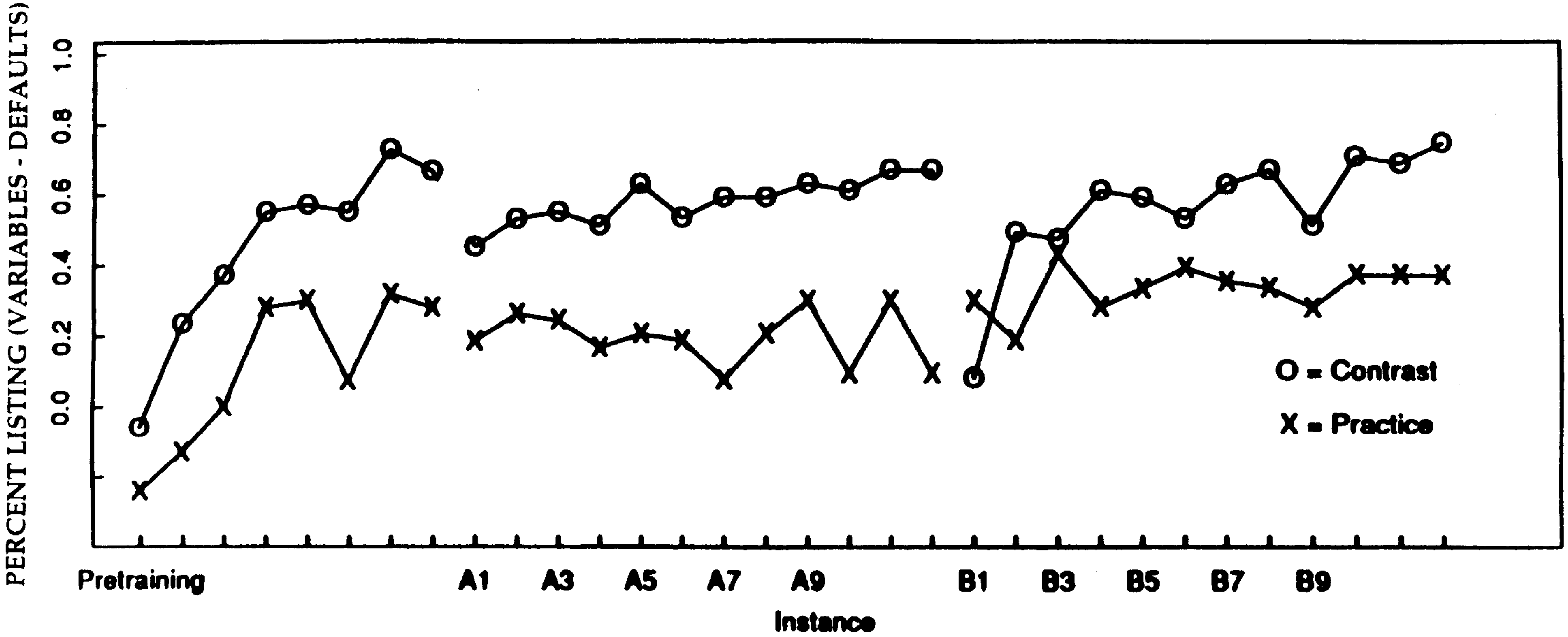
	ATTRIBUTES							
INSTANCES	1	2	3	4	5	6	7	8
1	2	1	2	1	1	2	1	1
2	1	2	1	2	2	2	2	1
3	2	2	2	1	2	2	2	1
4	1	1	2	2	1	1	1	1
5	1	2	1	2	2	1	2	2
6	1	1	2	2	1	2	1	1
7	2	1	1	1	1	1	1	2
8	2	2	1	1	2	1	2	2
9	1	1	1	2	1	2	1	2
10	1	2	2	1	2	1	2	1

MIXED SEQUENCE

	ATTRIBUTES							
INSTANCES	1	2	3	4	5	6	7	8
1	2	1	2	1	1	2	1	1
2	1	1	2	2	1	1	1	1
3	1	1	2	2	1	2	1	1
4	2	1	1	1	1	1	1	2
5	1	1	1	2	1	2	1	2
------------------	----	-----	-----	-----	-----	-----	-----	-----
6	1	2	1	2	2	2	2	1
7	2	2	2	1	2	2	2	1
8	1	2	1	2	2	1	2	2
9	2	2	1	1	2	1	2	2
10	1	2	2	1	2	1	2	1

BLOCKED SEQUENCE

10


PERCENT LISTING (VARIABLES - DEFAULTS)
1.0
0.8
0.6
0.4
0.2
0.0
Pretraining
A1
A3
A5
A7
A9
B1
B3
B5
B7
B9
Instance
O = Contrast
X = Practice

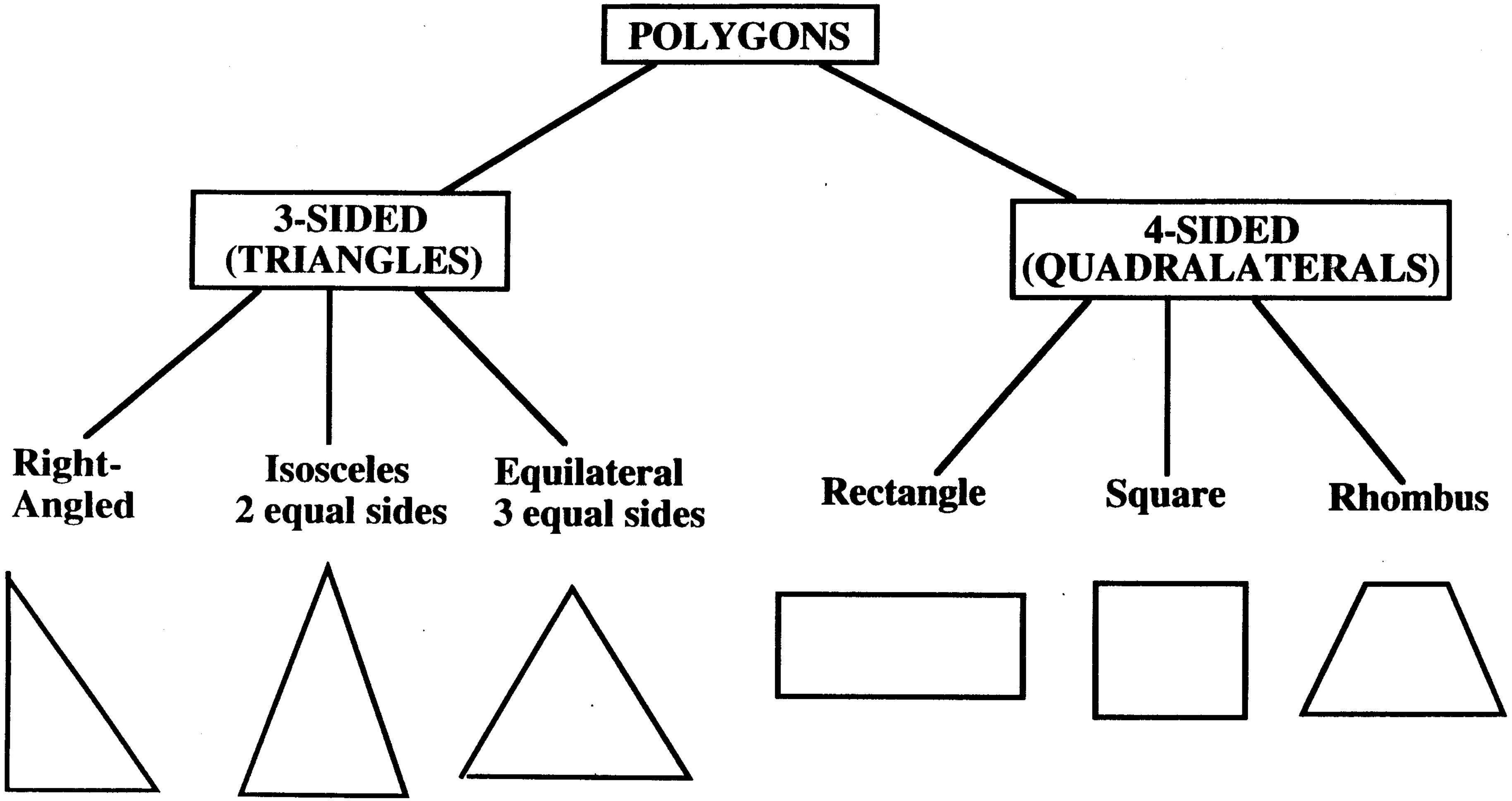
This difference in learning between blocked vs. mixed series is very interesting theoretically because the result contradicts most of the standard theories of unsupervised learning. For example, many clustering algorithms expect best performance when the model sees many contrasting examples in alternation rather than a block of one type. And connectionist models that use autoassociation to learn interfeature correlations predict either no difference due to sequencing stimuli or predict catastrophic retroactive interference with the blocked sequence.

The results instead point to a discrete process by which subjects invent a category, then acquire norms or defaults within that category that are sufficiently stable that subjects are able to be surprised by an unexpectedly large departure from those norms when they encounter the first instance of the alternate category. That surprise leads them to set up a new category and to begin learning its norms in a manner segregated from the norms learned about the first category. We've developed a simulation model that does just that, but I won't bother presenting it in this setting.

So those are the results I wished to present. The additional question for this conference is whether there are any practical applications of the result. I suppose some applications could certainly arise in educational settings where students are learning to distinguish between members of different categories, such as different animals, plants, flowers, airplanes, ships, or styles of residential architecture. In these cases, we have some chance of describing objects in terms of lists of features. Another example might be for students learning to identify geometric figures, such as regular polygons, as shown here (Overhead #11). Polygons can be divided according to their number of sides, with names provided for some that have special features. Thus, triangles can be classified as right triangles, isosceles, or equilateral depending on special features. If we wanted students to discover these classes for themselves, we could show them example triangles arranged in an order that was either blocked or random across the subcategories. Presumably they would discover the classes more quickly if they observed the examples in a blocked fashion. I think the blocking strategy would also work well with acquiring expertise in wine tasting: I can imagine that prospective wine tasters would learn to discriminate the wines more quickly if they tasted a collection of chardonnays, then rieslings, then sauvignon blancs in blocked fashion rather than tasting them in a mixed up sequence. You will all have the opportunity to test out this prediction at the receptions during this conference. [Overhead off]

Another application of the blocking idea would be general advice to unsupervised learners for when they explore an uncharted domain, especially if they have some control over the order in which they see examples. The advice is to avoid covering too much territory too quickly lest you get overwhelmed by the variability. Rather, it is better to start out slowly by exploring only relatively

11 II


POLYGONS
3-SIDED (TRIANGLES)
4-SIDED (QUADRALATERALS)
Right-Angled
Isosceles 2 equal sides
Equilateral 3 equal sides
Rectangle
Square
Rhombus

small variations of aspects of a given type of object--say, types of leaves on plants. In this way, you can learn a confident set of norms for that one type. Thereafter, using that as a firm foothold for classifying the domain, you can seek out a large contrasting type to learn next, and begin exploring small variations around that contrasting class. Our results suggest that, when it can be implemented, this strategy should produce fairly rapid discovery learning. That is, at least, one of the practical lessons I draw from this otherwise theoretical result.

Thank you for your attention.